

CRYSTALLOGRAPHIC STATISTICS

Progress and Problems


Edited by

S. Ramaseshan

M. F. Richardson

A. J. C. Wilson

Indian Academy of Sciences
Bangalore



Digitized by the Internet Archive
in 2018 with funding from
Public.Resource.Org

CRYSTALLOGRAPHIC STATISTICS

CRYSTALLOGRAPHIC STATISTICS

PROGRESS AND PROBLEMS

Edited by

S. Ramaseshan
M. F. Richardson
A. J. C. Wilson



INDIAN ACADEMY OF SCIENCES
BANGALORE 560 080

© 1982 by the Indian Academy of Sciences

Edited by S. Ramaseshan, M. F. Richardson
and A. J. C. Wilson and printed for
the Indian Academy of Sciences by
Macmillan India Press, Madras 600 002,
India

Table of Contents

Introduction		1
Introductory Lectures		
Crystallographic Statistics—General Review	HERBERT HAUPTMAN	5
Bayesian Statistics: An Overview	SIMON FRENCH AND STUART OATLEY	19
Intensity Statistics: Survey, Computer Simulation and the Heavy-Atom Problem	URI SHMUELI	53
Intensity Statistics		
Intensity Statistics: Non-Ideal Distributions in Theory and Practice	URI SHMUELI AND A. J. C. WILSON	83
Intensity Statistics and Probability of Validity of Phase Relations	G. B. MITRA AND SIKHA GHOSH	99
Effects of Heavy Atoms and Symmetry on the Cumulative Distribution Function of Normalised Structure Amplitudes	G. D. NIGAM AND SIKHA GHOSH	117
Measurability of Bijvoet Differences	S. PARTHASARATHY	133
Intensity Statistics and Non-Independence	A. J. C. WILSON	175
Statistics of Recorded Counts	J. L. DE BOER	179
Alternatives to R Tests	STUART M. ROTHSTEIN	187
The Residual Function R_2 as Discriminator Criterion in Structure Determination	A. T. H. LENSTRA	195

Variations on Least Squares

Alternatives to Least Squares	A. J. C. WILSON	225
A Robust/Resistant Technique for Crystal-Structure Refinement	W. L. NICHOLSON, E. PRINCE, J. BUCHANAN AND P. TUCKER	229
Calculation of the Electron-Density Distribution with an Account of Statistical Errors in Structure Amplitudes and Series Termination	A. A. SHEVYREV AND V. I. SIMONOV	265
On Data Reduction and Error Analysis for Single-Crystal Diffraction Intensities	ROBERT H. BLESSING AND GEORGE T. DETITTA	267
On the Problem of Secondary 'Least-Squares' Minima	R. ROTELBAUER	269
Wiener Methods		
Wiener Methods for Electron Density	D. M. COLLINS AND M. C. MAHAR	273
Subject Index		301
Author Index		309

Introduction

This volume had its origin in a 'Microsymposium' held on 22 August 1981, in the course of the Twelfth International Congress of the International Union of Crystallography. Certain authors were invited to review specific fields within the general area of crystallographic statistics, and the rest of the time allotted to the symposium was filled with selected contributed papers. Additional contributed papers were presented as posters at other times during the Congress, and papers of all three types were discussed at an *ad-hoc* session following the symposium. The Indian Academy of Sciences, through Professor S. Ramaseshan, has provided facilities for publication, and authors have provided manuscripts and read proofs within a rather tight time schedule. I and the co-chairman of the symposium, Professor Mary F. Richardson, are greatly indebted to Professor Ramaseshan and the authors.

Crystallographic statistics, in the sense intended in the title of this book, began almost by accident. In 1942, Professor S. H. Yü, happily able to be present at the Congress, submitted to *Nature* a paper on the determination of absolute from relative X-ray intensities. The Editors of *Nature* sent the paper to the Cavendish Laboratory for an opinion on its merit. The method was complicated and depended on the use of a set of tables not available in the United Kingdom, but Henry Lipson and I recommended its publication (Yü, 1942). The proposal set us arguing over a practicable method of achieving the same purpose, and the idea gradually emerged that the general

C.S.—1

level of the intensities of the various reflexions from a crystal must depend on the content at the unit cell and not on the details of the atomic arrangement. Lipson (unpublished) suggested calculating the structure factors for an arbitrary arrangement of the atoms in the unit cell and comparing the average calculated value with the average observed value for suitable groups of reflexions, but I wanted a tidier approach. I had some elementary statistics in mind in connexion with diffraction by disordered structures like cobalt and the copper-gold alloy AuCu_3 , and it soon became evident to me that the appropriate statistical variables to use were the X-ray intensities, not the structure factors. A very short calculation showed that the mean value of the intensity expressed in units of (electrons)² is equal to the sum of the squares of the scattering factors of all the atoms in the unit cell. Once obtained, this relation is practically obvious from conservation of energy, but it is, so far as I know, the first published result in the field now known as crystallographic statistics. My letter to *Nature* (Wilson, 1942), in effect a referee's report, has since become my most cited publication (Garfield, 1974, 1976), but it attracted no notice at the time, for very understandable reasons.

The subject remained quiescent until 1948, when Harker (1948) and Hughes independently rediscovered the main result of my 1942 paper. Hughes (1949) went further, and showed empirically that the distribution of the magnitudes of the structure factors was approximately normal. He gave four examples, all of centrosymmetric crystals, and it is not clear whether he realized that non-centrosymmetric crystals would have a different distribution of structure factors. Wilson (1949), using the central-limit theorem, derived the ideal distribution functions for both centrosymmetric and non-centrosymmetric crystals. The derivation rested on the explicit assumptions that the unit cell contained a sufficiently large number of atoms, that no one atom or small group of atoms dominated the scattering, and that the effect of other symmetry

elements, except centring, could be neglected. There was also an implicit assumption of negligible dispersion. This time the results were quickly taken up by other workers, both on the purely statistical side, and as the basis of direct methods of structure determination. The current President of the International Union of Crystallography, Professor Jerome Karle, and his co-worker Professor Herbert Hauptman, author of the introductory paper of the Microsymposium and of this volume, were the pioneers in the development of direct methods of an overtly statistical nature (Karle & Hauptman, 1953; Hauptman & Karle, 1953); there was, of course, a parallel development of non-statistical or covertly statistical direct methods, such as those of Sayre (1952) and Cochran (1952). A landmark in crystallographic statistics was the publication of the monograph by Srinivasan and Parthasarathy (1976). This, and such books as Giacovazzo (1980) on direct methods, may be consulted for the history of the growth of the subjects.

The present volume, like the symposium out of which it arose, was planned to concentrate on crystallographic statistics as such, and makes no attempt to include methods of structure determination, though it may be noted in passing that they were not neglected at the Ottawa Congress. It was intended to include papers on the following themes, in addition to Professor Hauptman's general introduction: Bayesian statistics, intensity statistics, statistics of recorded counts, alternatives to least squares, and Wiener methods for electron density. Ready consent was obtained from the speakers invited for four of these topics, but none of those invited to review alternatives to least squares was able to accept. Some contributed papers touching on the subject *via* altered weights are included after an editorial background note.

The full title of the symposium was 'Progress and Problems in Crystallographic Statistics'. The authors both of the invited and of the contributed papers have naturally concentrated on the 'progress', so the

balance has been redressed by editorial mention of some of the 'problems': bias in the estimation of parameters, in the note on alternatives to least squares; and doubts about the effect of correlation of atomic positions on the expressions for the probability distribution of intensities (p. 175 below). There is a further problem about intensity statistics, perhaps of academic interest only: the functional form of the distribution for really large intensities is as yet unknown (Wilson, 1980).

A. J. C. WILSON

References

- COCHRAN, W. (1952). *Acta Cryst.* **5**, 65–67.
 GARFIELD, E. (1974). *Curr. Contents*, 6 March; reprinted in Garfield (1976).
 GARFIELD, E. (1976). Pp. 37–44 in *Essays of an Information Scientist*, Vol. 2, Philadelphia: ISI Press.
 GIACOVAZZO, C. (1980). *Direct Methods in Crystallography*. London: Academic Press.
 HARKER, D. (1948). *Am. Mineral.* **33**, 764–765.
 HAUPTMAN, H. & KARLE, J. (1953). *Acta Cryst.* **6**, 136–141.
 HUGHES, E. W. (1949). *Acta Cryst.* **2**, 34–37.
 KARLE, J. & HAUPTMAN, H. (1953). *Acta Cryst.* **6**, 131–135.
 SAYRE, D. (1952). *Acta Cryst.* **5**, 60–65.
 SRINIVASAN, R. & PARTHASARATHY, S. (1976). *Some Statistical Applications in X-ray Crystallography*. Oxford: Pergamon Press.
 WILSON, A. J. C. (1942). *Nature (London)*, **150**, 151, 152.
 WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
 WILSON, A. J. C. (1980). *Technometrics* **22**, 629–630.
 YÜ, S. H. (1942) *Nature (London)*, **150**, 151, 152.

Crystallographic Statistics—General Review

BY HERBERT HAUPTMAN

*Medical Foundation of Buffalo, Inc., 73 High Street,
Buffalo, NY 14203, USA*

Abstract

The applications of statistical methods in crystallography fall into two major classes. The first is concerned with the study of the statistical properties of the intensities of X-rays diffracted by a crystal; the second with those of groups of related intensities. The latter is the basis for the analysis of the phase problem of X-ray crystallography by probabilistic methods, and an extensive literature describing these methods exists. However, this work is outside the scope of the present paper, which is concerned only with an overview of the statistical properties of the intensities of X-rays scattered by a crystal.

1. The basic Wilson distributions

Possibly the best place to start is with the very first applications of statistical methods in crystallography made by Wilson in 1949. Not only does this work illustrate in the clearest possible way the ideas of random variable and the probability distribution of a random variable, but it also makes an easily understood and important application of these ideas.

Imagine that a crystal structure in the space group $P1$ is given, that is to say a fixed set of atomic position vectors $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ is specified. Then the equation

$$E_{\mathbf{H}} = \frac{1}{\sigma_2^{1/2}} \sum_{j=1}^N f_j \exp (2\pi i \mathbf{H} \cdot \mathbf{r}_j), \quad (1)$$

where

$$\sigma_2 = \sum_{j=1}^N f_j^2, \quad (2)$$

and f_j is the scattering factor of the atom labelled j , defines the normalized structure factor $E_{\mathbf{H}}$ as a function of the reciprocal lattice vector \mathbf{H} . Clearly $E_{\mathbf{H}}$ is a complex-valued function of \mathbf{H} . By means of

$$|E_{\mathbf{H}}| = \frac{1}{\sigma_2^{1/2}} \left| \sum_{j=1}^N f_j \exp (2\pi i \mathbf{H} \cdot \mathbf{r}_j) \right|, \quad (3)$$

one obtains a real-valued (in fact non-negative-valued) function of \mathbf{H} .

For each crystal structure (3) defines a function of the reciprocal lattice vector \mathbf{H} so that there exists in fact an infinity of such functions. Thus (3) presents us with a class of functions of infinite and bewildering variety. How are we to make any sense of, or bring any order into, this hopelessly complex family of functions? In fact, can any general statement whatsoever be made about this large and varied class of functions?

In order to answer these questions we associate with each equation (3) another function which defines the distribution of values of the function $|E_{\mathbf{H}}|$. To clarify the notion of distribution of values, we ask, for example, what fraction of the values of $|E_{\mathbf{H}}|$ lies in the interval 0.0 to 0.1, or in the interval 0.1 to 0.2, or, more generally in any interval (a, b) where $0 \leq a < b$? The answer is given by the remarkably simple expression

$$\exp(-a^2) - \exp(-b^2), \quad (4)$$

or, as the mathematicians prefer, by

$$\int_a^b P(R) dR, \quad (5)$$

where

$$\left. \begin{aligned} P(R) &= 2R \exp(-R^2) \text{ if } R \geq 0, \\ P(R) &= 0 \text{ if } R < 0. \end{aligned} \right\} \quad (6)$$

The graph of $P(R)$ is shown in Fig. 1. Thus, the non-negative valued function, $P(R)$, of the real variable R defines the distribution of values of the function $|E_H|$ [equation (3)].

In this way we have arrived at the notion of a random variable (here the magnitude of a structure factor, $|E_H|$) and the probability distribution of a random variable [$P(R)$ in this case]. Now it is a remarkable fact that, under rather mild restrictions, $P(R)$ is independent of the crystal structure [although the function $|E_H|$, equation (3), clearly is not]. We say also that the probability that the random variable $|E_H|$ lie in the interval (a, b) is given by (4) [or (5)].

Fig. 1 clearly shows that the values of $|E_H|$ tend to cluster around 0.7, weak or absent reflections are very rare, and values of $|E_H|$ in excess say of 3 are also rare. This qualitative statement can be made more precise, as shown next.

With the function $P(R)$ in our possession it is a relatively simple matter, at least in principle, to answer certain questions concerned with the distribution of values of the function $|E_H|$. For example, the average value of $|E_H|^2$ and the variance of $|E_H|^2$ are easily

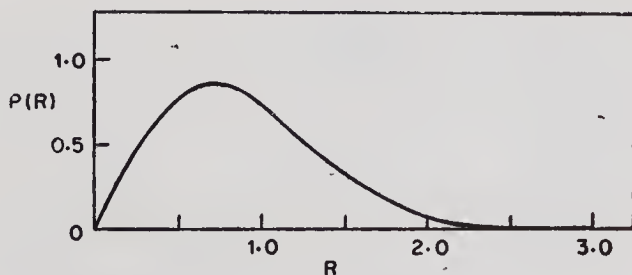


Fig. 1. The probability distribution $P(R)$, equation (6) of the magnitude of a structure factor $|E_H|$, in Pl.

found (at least in this case) by means of the formulas

$$\begin{aligned} \overline{|E_{\mathbf{H}}|^2} &= \int_{-\infty}^{\infty} R^2 P(R) dR \\ &= \int_0^{\infty} 2R^3 \exp(-R^2) dR = 1, \quad (7) \end{aligned}$$

$$\begin{aligned} \text{Var}(|E_{\mathbf{H}}|^2) &= \langle (|E_{\mathbf{H}}|^2 - \overline{|E_{\mathbf{H}}|^2})^2 \rangle \\ &= \int_{-\infty}^{\infty} (R^2 - 1)^2 P(R) dR = 1. \quad (8) \end{aligned}$$

If the space group is $P\bar{1}$ then $P(R)$ is Gaussian (Fig. 2):

$$\left. \begin{aligned} P(R) &= \sqrt{\frac{2}{\pi}} \exp\left(-\frac{1}{2} R^2\right) \text{ if } R \geq 0, \\ P(R) &= 0 \text{ if } R < 0, \end{aligned} \right\} \quad (9)$$

which should be compared with (6). Fig. 2 shows, in sharp contrast to Fig. 1 for space group $P1$, that weak or absent reflections are now relatively numerous and that very strong reflections, while rare, occur with greater frequency in $P\bar{1}$ than in $P1$. Although, as it turns out, the average value of $|E_{\mathbf{H}}|^2$ is unity for both space groups, even a superficial inspection of Figs. 1 and 2 shows that the dispersion of values of $|E_{\mathbf{H}}|^2$ about its mean is greater for $P\bar{1}$ than for $P1$. The quantitative result is given next.

The further comparison of (6) and (9) suggests in fact a method for distinguishing between $P1$ and $P\bar{1}$ by a purely statistical analysis of the distribution of values of the magnitudes of the observed structure

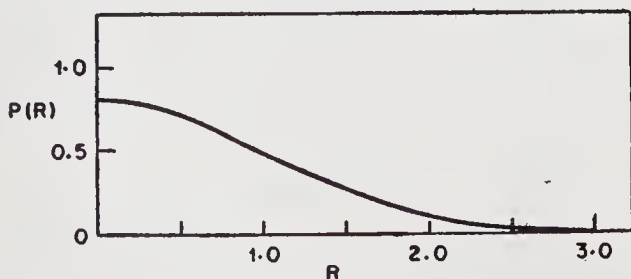


Fig. 2. The probability distribution $P(R)$, equation (9), of the magnitude of a structure factor $|E_{\mathbf{H}}|$, in $P\bar{1}$.

factors. For example, the mean and variance are now given by

$$\overline{|E_{\mathbf{H}}|^2} = \int_0^\infty \sqrt{\frac{2}{\pi}} R^2 \exp\left(-\frac{1}{2} R^2\right) dR = 1, \quad (10)$$

$$\begin{aligned} \text{Var}(|E_{\mathbf{H}}|^2) = \int_0^\infty \sqrt{\frac{2}{\pi}} (R^2 - 1)^2 \\ \exp\left(-\frac{1}{2} R^2\right) dR = 2, \quad (11) \end{aligned}$$

so that the comparison of (8) and (11) serves to distinguish the two space groups. As anticipated, the variance, a measure of dispersion about the mean, is in fact greater for $P\bar{1}$ than for $P1$.

By means of a more detailed study of the effects of the various symmetry elements on these distributions, together with the observation of systematic absences, one arrives at a method for determining the space group in which the statistical analysis of the observed intensities plays the major role.

It should be noted that, although the probability distributions $P(R)$ [(6) and (9)] describe the distribution of values of the magnitude of a structure factor, a simple change of variable enables one to replace these distributions by new ones describing the distribution of values of the intensity I of a reflection.

2. Estimating the values of weak or unobserved reflections

Expressions for the probability $p_i(I_0 | I)$ that a reflection of true intensity I will have an observed value I_0 (possibly negative) have been found for different counting modes, specifically for fixed-time counting (*i.e.* the number of counts per specified length of time interval), for fixed-count timing (*i.e.* the counting rate for specified number of counts) and for variations of these (Wilson, 1980). These distributions serve as the basis for a proper statistical method for estimating

the values and standard deviations of the intensities of measured reflections, particularly in the important case that the intensities are weak or measured to be negative.

The integrated intensity of a reflection is obtained as the difference between a counting rate averaged over a region of reciprocal space which includes the reflected intensity and that averaged over a nearby region which excludes the reflected intensity. Owing to statistical fluctuations in the counting rates, a measured intensity may be negative, although the true intensity must of course be non-negative. Can one obtain an improved estimate of the true intensity and its standard deviation by taking into account its known *a priori* probability distribution?

The Bayesian approach to this problem (French & Wilson, 1978) interprets probability distributions as degrees of belief in the possible values of the intensity rather than the distribution of values of the intensity. Prior to measuring an intensity we have a certain distribution of belief in its possible values and this distribution is changed in a known way after the measurement is made. Thus, by calculating the *a posteriori* expectation value and variance one obtains an improved estimate of the true intensity and its variance from the measured values which is of particular importance in the case that the measured intensity is weak or negative.

More specifically, if one denotes by $p(I)$ the probability distribution of the intensity of a reflection [derived *e.g.*, from (6) or (9)], by $p_i(I_0 | I)$ the (*a priori*) conditional probability distribution of the observed intensity I_0 of a reflection, given that its true value is I , and by $p_f(I | I_0)$ the (*a posteriori*) conditional probability distribution of the intensity I of a reflection, given that the observation I_0 has been made, then Bayes' theorem (see, *e.g.*, pages 108–114 of Feller, 1960) states that

$$P_f(I | I_0) = K p_i(I_0 | I) p(I), \quad (12)$$

where K is a suitable scaling parameter. From (12) one may calculate the *a posteriori* expected value and variance of the intensity of a reflection after the observation I_0 (possibly negative) has been made.

It is important to derive the best estimates possible for observed intensities and their reliability, particularly when these are weak or unobserved, for several reasons. First, in this way one reduces or eliminates the bias in the resulting structural parameters. Next, improved estimates of the structure factor moduli and their associated errors are also needed, for example, in the calculation of Fourier series or difference syntheses. Again, if a large number of reflections are unobserved because they are weak, the procedure may serve an important function in the determination of the crystal structure by direct methods which usually require a great over-redundancy of data for success and in which the weak intensities play an important role. Finally, in view of recent developments in integrating the techniques of direct methods with isomorphous replacement and anomalous dispersion, it is likely that the reflections of weak or moderate intensity will play an increasingly important role in the solution of the phase problem and in structure determination, particularly in the macromolecular case (Hauptman, 1981*a, b*).

3. Structural isomorphism

For an isomorphous pair of structures normalized structure factors $E_{\mathbf{H}}$ and $G_{\mathbf{H}}$ are defined by

$$E_{\mathbf{H}} = \frac{1}{\sigma_2^{1/2}} \sum_{j=1}^N f_j \exp (2\pi i \mathbf{H} \cdot \mathbf{r}_j),$$

$$\sigma_2 = \sum_{j=1}^N f_j^2, \quad (13)$$

$$G_{\mathbf{H}} = \frac{1}{\tau_2^{1/2}} \sum_{j=1}^N g_j \exp (2\pi i \mathbf{H} \cdot \mathbf{r}_j),$$

$$\tau_2 = \sum_{j=1}^N g_j^2. \quad (14)$$

We seek the joint probability distribution of the pair of magnitudes $|E_{\mathbf{H}}|$, $|G_{\mathbf{H}}|$, *i.e.* the function $P(R, S)$ which defines the distribution of values of the ordered pair of non-negative real numbers $(|E_{\mathbf{H}}|, |G_{\mathbf{H}}|)$, in much the same way that $P(R)$ defines the distribution of values of $|E_{\mathbf{H}}|$ alone. Thus the fraction of pairs $(|E_{\mathbf{H}}|, |G_{\mathbf{H}}|)$ for which $|E_{\mathbf{H}}|$ lies in the interval (a, b) and $|G_{\mathbf{H}}|$ lies in the interval (c, d) is given by the double integral.

$$\int_{R=a}^b \int_{S=c}^d P(R, S) dR dS, \quad (15)$$

where, as it turns out, the function $P(R, S)$ is defined by

$$\left. \begin{aligned} P(R, S) &= \frac{4RS}{1-a^2} \exp \left\{ -\frac{R^2 + S^2}{1-a^2} \right\} \\ I_0 \left(\frac{2aRS}{1-a^2} \right) &\quad \text{if } R \geq 0 \text{ and } S \geq 0, \\ P(R, S) &= 0 \quad \text{if } R < 0 \text{ or } S < 0, \end{aligned} \right\} \quad (16)$$

I_0 is the Modified Bessel Function (Fig. 3), and

$$a^2 = \frac{(\sum_{j=1}^N f_j g_j)^2}{(\sum_{j=1}^N f_j^2)(\sum_{j=1}^N g_j^2)}. \quad (17)$$

$P(R, S)$ is said to be the joint probability distribution of the pair of random variables $(|E_{\mathbf{H}}|, |G_{\mathbf{H}}|)$.

The graph of the Bessel function $I_0(x)$ shows that, rather like $\cosh x$, this function grows at approximately an exponential rate as x tends toward $\pm\infty$.

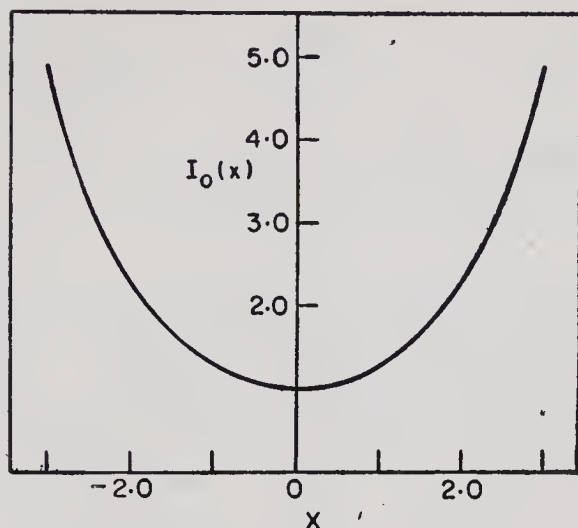


Fig. 3. The modified Bessel function, $I_0(x)$.

Nevertheless, owing to the presence of the descending exponential factor in (16), $P(R, S)$, which is equal to zero when $R=0$ or when $S=0$, again tends toward zero with increasing R or S , but is of course positive for intermediate values of R and S .

Equation (16) enables one to calculate the correlation coefficient r of the pair $(|E_H|^2, |G_H|^2)$. (See Watson, 1958, for the needed integral formulas.)

$$r = \frac{\text{Cov}(|E_H|^2, |G_H|^2)}{\{\text{Var}(|E_H|^2)\}^{1/2} \{\text{Var}(|G_H|^2)\}^{1/2}}. \quad (18)$$

$$r = \frac{\langle (|E_H|^2 - \overline{|E_H|^2}) (|G_H|^2 - \overline{|G_H|^2}) \rangle_H}{\langle (|E_H|^2 - \overline{|E_H|^2})^2 \rangle_H^{1/2} \langle (|G_H|^2 - \overline{|G_H|^2})^2 \rangle_H^{1/2}}, \quad (19)$$

$$\overline{|E_H|^2} = \overline{|G_H|^2} = \int_0^\infty \int_0^\infty R^2 P(R, S) dR dS = \int_0^\infty \int_0^\infty S^2 P(R, S) dR dS = 1, \quad (20)$$

$$\text{Cov}(|E_H|^2, |G_H|^2) = \int_0^\infty \int_0^\infty (R^2 - 1)(S^2 - 1) P(R, S) dR dS = \alpha^2, \quad (21)$$

$$\text{Var}(|E_{\mathbf{H}}|^2) = \text{Var}(|G_{\mathbf{H}}|^2) = \int_0^\infty \int_0^\infty (R^2 - 1)^2 P(R, S) dR dS = 1 \quad (22)$$

and, finally, from (18),

$$r = \alpha^2. \quad (23)$$

Inspection of (17) and (23) shows that in the extreme case that $f_j = g_j$ for every j then $|E_{\mathbf{H}}|^2 = |G_{\mathbf{H}}|^2$ for every \mathbf{H} , and the correlation coefficient of the pair $(|E_{\mathbf{H}}|^2, |G_{\mathbf{H}}|^2)$ is unity, as expected. In general, however, r is positive and less than unity. Clearly, in the case of perfect isomorphism, r is a positive constant as a function of $\sin\theta/\lambda$. In the case of imperfect isomorphism on the other hand, r is a monotonically decreasing function of $\sin\theta/\lambda$. Thus $P(R, S)$ leads to a method for determining the degree of isomorphism between two structures. In fact r , as a function of $\sin\theta/\lambda$, may be taken as a measure of the degree of isomorphism of the two structures, *i.e.* the degree to which the two structures coincide. Application to protein crystallography, for which the isomorphous replacement technique is the most important tool, is clear.

Finally, the case that the G -structure is a trial, or model structure, possibly incomplete, is also included. In this case a more detailed analysis leads to the joint probability distribution $P(R, S)$ dependent on the parameter $\langle |\Delta r| \rangle$, the average error of the trial structure. Then the average error is expressible in terms of the correlation coefficient r , although the normalized discrepancy index

$$R = \frac{\langle |E_{\mathbf{H}}| - |G_{\mathbf{H}}| \rangle_{\mathbf{H}}}{\langle |E_{\mathbf{H}}| \rangle_{\mathbf{H}}}, \quad (24)$$

is more commonly used. Clearly then this distribution plays an important role in the process of refinement of crystal structures. (See Srinivasan & Parthasarathy, 1976, for further details.)

4. Weighting

4.1. *Least-squares refinement*

The least-squares refinement process of crystal structures consists simply of the solution by least-squares of the highly redundant structure factor equations. In order to minimize systematic error in the determination of the structural parameters it is essential to weight these equations correctly and, to this end, estimates of the uncertainties in the observed intensities, along the lines briefly described earlier, are necessary.

4.2. *Fourier syntheses*

The electron density function ρ is represented by a Fourier series the coefficients of which are the structure factors

$$F = |F| \exp(i\phi). \quad (25)$$

In practice the magnitudes $|F|$ are obtained from experiment and are therefore subject to error the magnitudes of which may be estimated as described earlier, *e.g.* by taking into account fluctuations due to counting statistics, instrument instability, and inadequate correction factors and then employing the Bayesian approach (*via* the *a posteriori* probability distribution) to estimate the standard deviation. The phases ϕ are derived from the observed magnitudes $|F|$ and are therefore subject to additional errors the distribution of which depends on the nature of the phase determination process. Thus the structure factor F is a random variable, and its probability distribution determines how the Fourier coefficients are to be weighted in order to yield, in some sense, the 'best' electron density function ρ . Similar remarks apply, for example, to the use of the Patterson function to locate heavy atom positions employing data from a pair of isomorphous crystals.

No matter how the phases are determined they are subject to error. In the heavy atom method, for example, phases calculated from the heavy atom positions are clearly only an approximation to the true phases. Phases obtained by single or multiple isomorphous replacement are subject to error because of experimental error in the observed intensities as well as imperfect isomorphism. Phases determined by the method of anomalous dispersion or direct methods are subject to similar errors.

Since

$$F_{\text{true}} \neq F_{\text{obs}}, \quad (26)$$

it follows that

$$\rho_{\text{true}} \neq \rho_{\text{obs}}. \quad (27)$$

Define

$$\Delta\rho = \rho_{\text{true}} - \rho_{\text{obs}} \quad (28)$$

and the best Fourier that one which minimizes

$$\int_V (\Delta\rho)^2 dV. \quad (29)$$

This definition of the best Fourier then serves as the basis for the derivation of a suitable weighting function W . (See, *e.g.* Srinivasan & Parthasarathy, 1976, for further details.)

4.3. *Tangent formula*

The most widely used formula for expanding and refining a basis set of phases is the tangent formula:

$$\tan \phi_{\text{H}} = \frac{\langle |E_{\text{H}} E_{\text{H-K}}| \sin(\phi_{\text{K}} + \phi_{\text{H-K}}) \rangle_{\text{K}}}{\langle |E_{\text{K}} E_{\text{H-K}}| \cos(\phi_{\text{K}} + \phi_{\text{H-K}}) \rangle_{\text{K}}} = \frac{T}{B}, \quad (30)$$

where $\sin \phi_{\text{H}}$ has the same sign as T and $\cos \phi_{\text{H}}$ has the same sign as B . Although it may not be the best technique for this purpose, it does give good results, in general, and is most efficient. The stability of the formula is improved if the averages are weighted

by means of an increasing function of $T^2 + B^2$, the rationale for this procedure depending on a rather detailed study of the derivation, *via* probabilistic techniques, of the tangent formula. The formula, with suitable weights, has assumed greater importance in recent months with the unexpected discovery by Yao Jia-Xing (1981) that a randomly chosen basis set of phases will surprisingly often converge to the correct answer, so that the tangent formula alone, properly weighted, becomes an important tool for structure determination.

This research was supported by National Science Foundation Grant No. CHE79-11282 and Grant No. GM-26195 from the National Institute of General Medical Sciences, DHHS.

References

- FELLER, W. (1969). *An Introduction to Probability Theory and Its Applications*. John Wiley & Sons. Inc.
- FRENCH, S. & WILSON, K. (1978). *Acta Cryst.* A34, 517–525.
- HAUPTMAN, H. (1981a). *Am. Crystallogr. Assoc. Winter Meet.*, Abstract C4.
- HAUPTMAN, H. (1981b). XII Intern. Congr. of Crystallography, Abstract 17.2–09.
- JIA-XING, YAO (1981). *Acta Cryst.* A37, 642–644.
- SRINIVASAN, R. & PARTHASARATHY, S. (1976). *Some Statistical Applications in X-Ray Crystallography*. Pergamon Press.
- WATSON, G. N. (1958). *A Treatise on the Theory of Bessel Functions*. Cambridge University Press.
- WILSON, A. J. C. (1949). *Acta Cryst.* 2, 318–321.
- WILSON, A. J. C. (1980). *Acta Cryst.* A36, 929–936.

Bayesian Statistics: An Overview

BY SIMON FRENCH

*Department of Decision Theory, University of
Manchester, Manchester, M13 9PL, England*

AND STUART OATLEY

*Laboratory of Molecular Biophysics, Department of
Zoology, University of Oxford, South Parks Road,
Oxford, OX1 3PS, England*

Abstract

We describe the basic principles of the Bayesian approach to statistical analysis. We show how it leads to sensible estimates of structure factor moduli from intensity observations, whether the latter are positive or negative. For diffractometer data collected using a step-scan method, we develop a profile-fitting approach to primary data reduction based upon the Bayesian three-stage regression model. Finally, we indicate how a Bayesian approach to model choice may lead to a satisfactory alternative to Hamilton's *R*-test as a means of choosing between differing molecular structures that result from refinements to the same data but under different sets of soft constraints.

Introduction

It is a gross simplification to describe present day statistical thinking as divided between two schools: the Bayesian and the frequentist. Nonetheless, it is a simplification that we shall make; because by doing so we may introduce and emphasise the distinctive flavour of the Bayesian approach. The difference

between the two schools lies mainly in their interpretations of the concept of probability; so we shall begin there.

A frequentist holds that probability has meaning only as the numerical representation of variability actually present within a system. Because of this he cannot give a technical probabilistic meaning to questions such as the following.

- (i) What is the most probable value of a parameter?
- (ii) Within what range of values does an unknown parameter most probably lie?
- (iii) Is it improbable that a particular hypothesis is true?

The first pair of questions is meaningless to him because the parameters of a model do not vary: they are fixed even if they are unknown. The third question is meaningless because the truth or falsehood of a hypothesis is immutable. None of these questions refer to variability actually present within a system. However, while they may be meaningless to him within his technical language, they are the everyday expression of his motives in a statistical investigation. It is to answer such questions that he has developed his methods of estimation, confidence interval construction, and hypothesis testing.

This conflict between everyday language and the technical language of frequentist statistics may seem an esoteric matter for the philosopher; however, it does have importance for the practical scientist. The need to provide answers to questions that cannot be framed technically has resulted in frequentist statistics being based upon arguments that are exceedingly subtle: some would say, contorted. The necessary 'double-think' of the frequentist approach makes it particularly difficult for the scientist to learn, understand, and use the resulting methods of inference. How many students find their statistics courses easy?

More importantly, real experiments are seldom quite like the stereotypes found in statistical text-books. Thus to interpret an actual data set it is usually necessary to modify a standard technique or, perhaps, develop an entirely new one. The subtleties of argument required to complete these tasks are often beyond the scientist; and many data sets are poorly interpreted, information is lost, and, very occasionally, false conclusions drawn.

The Bayesian approach does not lead to this conflict of language. On the contrary, its technical language is a direct numerical representation of the scientist's everyday language. Here probability is taken to represent the various degrees of belief or uncertainty that a scientist has in the truth of propositions about the system under observation. For instance, the more likely that a parameter has the value 1.54 (say) the higher is the numerical probability of the proposition 'this parameter has the value 1.54'. All uncertainty is modelled through probability and, in particular, the three questions quoted above translate directly into the technical language of Bayesian statistics. As a consequence Bayesian argument is intuitive, easily learnt, and, most importantly, easily developed. There is no difficulty in constructing methods appropriate to each particular investigation.

Typically a Bayesian analysis proceeds as follows. The scientist first develops a physical model for the system under study. Within this model there will be many unknown parameters, perhaps even unknown functional forms; but nothing will be completely unknown. There may be theoretical reasons why a parameter must be positive. Previous investigations may limit the possible range of a parameter. Exact functional forms may be unknown, but they may be expected to possess properties of smoothness, symmetry or unimodality, etc. All such prior information may be modelled probabilistically; and it is much of the purpose of later sections of this paper to indicate how this may be done.

The next step is the design and execution of an experiment. Here the scientist must ask himself what he would expect to observe if he knew the unknown quantities exactly. He considers this question for each possible value or form of the unknowns, and from his answers describes the structure of his experiment probabilistically. Again we shall illustrate precisely how this may be done, later in the paper.

At this point in the analysis the scientist may sit back and let the basic laws of mathematics take over. For, once the results of the experiment have been observed, all that need be done is to apply Bayes' theorem (hence the name Bayesian statistics) in order to combine the prior information about the unknowns with that inherent in the experimental data. The resulting posterior distribution is the probabilistic representation of the synthesis of all the information that the scientist has; and it provides the answers to whatever questions might concern him. According to circumstances, the mean, mode, or median of the posterior distribution may serve as a suitable estimate of the parameters. It is straightforward to calculate the most probable range for a parameter; and hypothesis testing simply becomes a matter of comparing relative posterior probabilities.

Although we shall avoid as much mathematical notation as possible, instead concentrating on illustrating and, we hope, illuminating Bayesian methods, some will be necessary; and expressing the above paragraphs symbolically will serve as an excellent introduction. The scientist begins with a physical model, in which there are some unknown quantities, which we shall denote by θ . The prior information about θ is represented by the prior distribution, which we assume to have the probability density function $p_\theta(\cdot)$. Throughout this paper we shall use subscripts to indicate quantities about which beliefs are being expressed; this enables us to use a lower case p to denote all probability density functions.

The experiment gives rise to an observation \mathbf{X} ,

which the scientist expects to occur with probability density $p_X(\cdot | \theta)$. Note that this distribution is conditional on the unknowns θ . It describes the variation that the scientist would predict in his observations on the basis of his physical model if he knew the exact values of the unknowns θ .

If the observations $X = x$ are actually made, then Bayes' theorem gives the scientist's belief in θ updated by the experimental data as:

$$p_{\theta}(\theta | x) \propto_{\theta} p_X(x | \theta) \cdot p_{\theta}(\theta). \quad (1)$$

The \propto_{θ} means 'is proportional to as a function of θ '; x being the fixed, observed value. The constant of proportionality is determined by the condition that a probability density must integrate to unity. $p_X(x | \theta)$, when thought of as a function of θ with X fixed at the observed value x , is known as the likelihood function. Thus (1) may be remembered by

posterior \propto likelihood \times prior.

The structure of Bayesian analysis is summarised in Fig. 1.

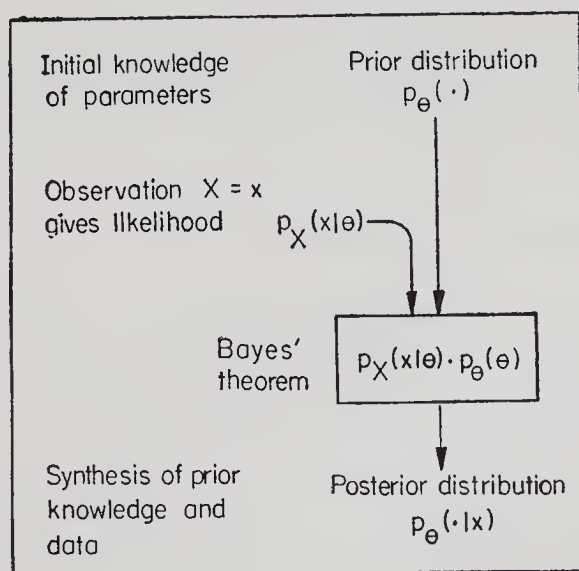


Fig. 1. The structure of Bayesian analysis.

It would be foolish to pretend that the Bayesian approach is without its critics, but we shall not discuss their objections here. The references cited below do that far more effectively than we could. However, we do indicate and briefly discuss one important objection in § 4 of this paper. Our purpose in the remaining sections is to introduce Bayesian analyses of some crystallographic problems. We shall avoid technical details, which are available elsewhere in the literature, and concentrate on providing an overview.

Barnett (1973) provides a very readable account of the differences between Bayesian and frequentist statistics. An excellent introduction to Bayesian analysis may be found in the early chapters of Box & Taio (1973); the later chapters are important for their technical details. Other books which develop the Bayesian approach are: Jeffreys (1961), Lindley (1965), DeGroot (1970), and De Finetti (1974). Box (1980) is particularly important for its setting of Bayesian statistics within the Scientific Method. Within the crystallographic literature few applications of Bayesian ideas have been reported; we know only of Mendes & De Polignac (1973), French & Wilson (1978), French (1978), and Oatley & French (1981). We might also refer to the excellent statement of the phase problem within a Bayesian framework in the first few paragraphs of Hauptman & Karle (1953).

1. Negative intensity observations

Reflections with small structure factor moduli have always led to difficulties. Their true intensities are, of course, non-negative, but their observed intensities may not be, because of counting statistics or photographic recording errors. When a measured intensity is positive, its square root forms a sensible estimate of the structure factor modulus. However, what should be done when the observation is negative? Various suggestions have been made: all negative

observations should be omitted from the data, *i.e.* treated as unobserved; they should be set to zero and included in the data; or they should be set to some arbitrary, constant fraction of the mean intensity in the data. Unfortunately none of these procedures is entirely satisfactory. All can lead to biases in the final structure. Moreover, they can lead to difficulty in interpreting Fourier, particularly difference Fourier, syntheses in the early stages of the structure determination. As shown in detail by French & Wilson (1978) and as sketched briefly below, a Bayesian analysis leads to a natural and straightforward solution to the problem.

At a given reflection the parameter that concerns us is the true intensity. It is arguable, particularly in the light of the preceding discussion, that the true structure factor modulus is the parameter of interest; but, as will become apparent, this would lead to precisely the same results. Thus the unknown parameter in our physical model is the true intensity, which we shall denote by J . In our notation of § 1, $\theta = J$. Our first task then is to consider our prior knowledge of J and so define the prior density $p_J(\cdot)$. The most obvious piece of information that we have is that J is non-negative; French & Wilson (1978) discuss a density which embodies this and only this information. However, as they point out, we invariably know rather more than just the sign of J . Taken as a whole, any moderate or large data set obeys Wilson's (1949) statistics. So, for an acentric reflection

$$p_J(J) = \begin{cases} (\Sigma)^{-1} \exp(-J/\Sigma) & \text{if } J \geq 0, \\ 0 & \text{otherwise;} \end{cases} \quad (2)$$

and for a centric reflection

$$p_J(J) = \begin{cases} (2\pi \Sigma J)^{-1/2} \exp(-J/2\Sigma) & \text{if } J \geq 0, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

For small molecules, where one may assume that the atoms are uniformly and independently distributed about the unit cell, a simple theoretical derivation shows that Σ is the sum of squares of the atomic scattering factors of all the atoms in the unit cells. For larger molecules the presence of secondary structure makes the assumption that the atoms are independently distributed untenable. In such cases French & Wilson conjectured and showed empirically that Σ may be taken to be the mean intensity in the appropriate shell of reciprocal space. Wilson (1981) has recently provided a theoretical justification of this conjecture.

It would, of course, be possible to use prior densities specific to space groups of higher symmetry, but to our knowledge this has not been done.

With the prior distribution now defined we turn to the experimental observation. We shall denote the observed intensity by I . By 'observed intensity' we mean the following. We assume that all relevant data sets, collected either by diffractometer or photographic methods, have been corrected for Lorentz, polarisation, absorption, extinction, and radiation-damage effects, have been reduced to a common scale, and have been merged over equivalents. I is this 'merged intensity' containing all the available observational information at the given unique reflection. All the operations needed to produce this merged intensity are assumed to have been carried out on the raw intensity measurements, be they positive or negative.

I forms our experimental observation; so in the notation of § 1 we have $\mathbf{X} = I$. Hence we must now consider the probability density $p_I(\cdot | J)$ (i.e. $p_{\mathbf{X}}(\cdot | \theta)$ in § 1). Throughout we shall assume that $p_I(\cdot | J)$ is a normal density, viz.

$$I \sim N(J, \sigma^2). \quad (4)$$

Thus, aside from normality, we are also assuming that I is an unbiased observation on J with known

variance σ^2 . These assumptions are discussed by French & Wilson (1978).

The posterior distribution for J is now given by Bayes' theorem (*cf.* (1)):

$$p_J(J|I) \propto p_I(I|J) \cdot p_J(J). \quad (5)$$

Note that $p_J(J|I) = 0$ for $J < 0$ because $p_J(J)$ occurs multiplicatively in (5) and is zero for this range of J (see (2) and (3)). Thus our prior knowledge that J must be non-negative is carried through to the posterior distribution.

Most, if not all, crystallographic structure solutions do not use the posterior density for the intensity in its entirety; but use approximations based upon its mean and variance, or upon the mean and variance of its square root, the structure factor modulus. Least squares refinement is an extremely common example of a solution technique that requires just these. So usually we do not need the full density $p_J(J|I)$, but only its moments:

$$E_J(J|I) = \int J \cdot p_J(J|I) \cdot dJ, \quad (6)$$

$$\text{Var}_J(J|I) = \int (J - E_J(J|I))^2 \cdot p_J(J|I) \cdot dJ, \quad (7)$$

or, letting $F = \sqrt{J}$,

$$E_J(F|I) = \int F \cdot p_J(J|I) \cdot dJ, \quad (8)$$

$$\text{Var}_J(F|I) = \int (F - E_J(F|I))^2 \cdot p_J(J|I) \cdot dJ. \quad (9)$$

Earlier we stated that it did not matter whether we took the true intensity or the true structure factor modulus as the parameter of interest. Expressions (8) and (9) confirm that by taking J as the parameter we do not lose the ability to estimate F and to give a guide to the precision of this estimate. Indeed, had we taken F as the parameter, our analysis would have led to expressions completely equivalent to (6)–(9). Our prior distribution $p_F(\cdot)$ would have been derived from Wilson's statistics for the structure factor

modulus and would thus have been (2) or (3) with the change of variable $F = \sqrt{J}$ and the introduction of the appropriate Jacobian $|\partial J/\partial F|$. The Jacobian would have remained throughout the resulting analysis, ensuring that it was identical to that above but with a simple change of variable.

Expressions (6) – (9) may look horrendous, but they are nonetheless relatively easy to evaluate or, at least, approximate. French & Wilson (1978) give details.

It should be noted that this Bayesian analysis applies to all reflections, whatever the observed intensities. There are no *ad-hoc* cut-off points with positive observations treated one way and negative another. This consistency of treatment is typical of the Bayesian approach and adds much to its intuitive appeal.

In Fig. 2 we summarise and illustrate our analysis. Lewis (1981) has recently extended and developed these ideas to include anomalous scattering information. He was interested not in estimating the intensities themselves, but rather anomalous differences across Bijvoet pairs. Using appropriate prior distributions drawn from Srinivasan & Parthasarathy (1976) and taking the differences in measured intensities across Bijvoet pairs as his observations, he has produced sensible estimates of anomalous differences and from these successfully located the heavy atoms in an isomorphous protein derivative. Without the Bayesian method these atoms had proved difficult to locate.

2. The Bayesian three-stage model

No physical model perfectly explains data, however free they are from experimental error. There is always some discrepancy between the mathematical behaviour of the model and the actual behaviour of the system being modelled. Usually this modelling error is several orders of magnitude smaller than the experimental

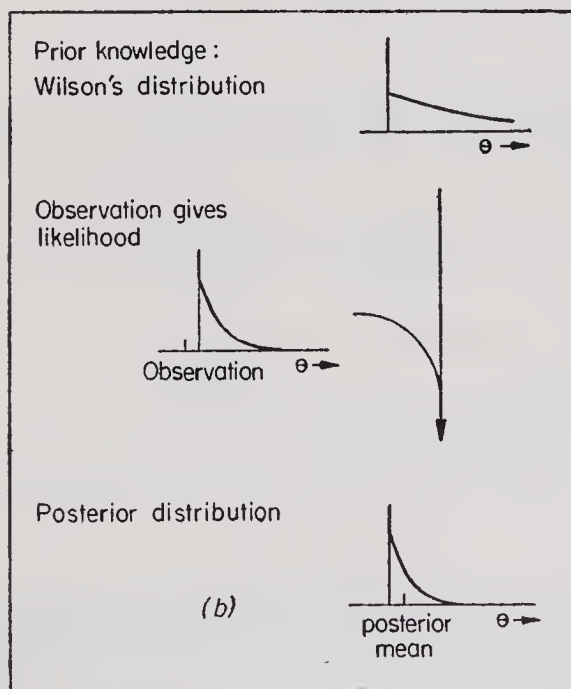
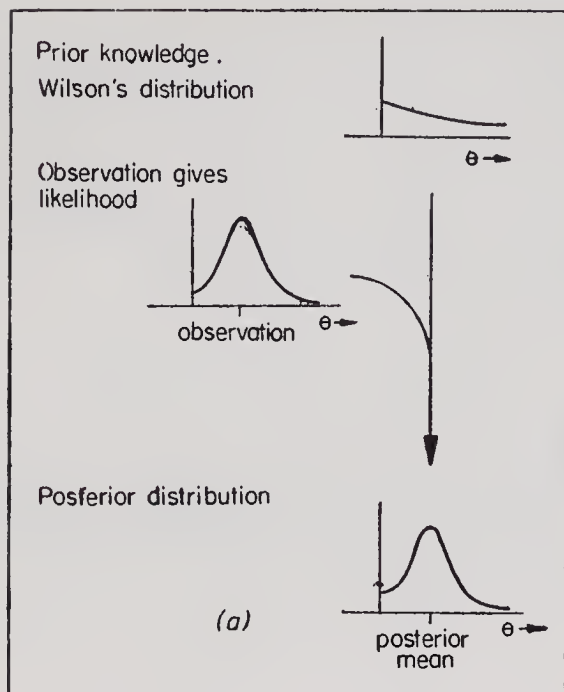


Fig. 2. Illustration of the Bayesian analysis: (a) when the observation is positive; (b) when the observation is negative.

error present, and thus can safely be ignored. However, there are occasions when the modelling error is sufficiently large to require special treatment in the statistical analysis. In this section we explain how modelling error can be included in a Bayesian analysis. To do this it will be necessary to return to (1) and discuss Bayes' theorem a little further. Bayes' theorem,

$$p_{\theta}(\theta | \mathbf{x}) \propto_{\theta} p_{\mathbf{X}}(\mathbf{x} | \theta) \cdot p_{\theta}(\theta), \quad (10)$$

is really little more than the re-expression of the joint distribution of \mathbf{X} and θ . By the definition of conditional probability densities (ignoring the niceties of measure, *i.e.* integration theory) we have:

$$\begin{aligned} p_{\theta}(\theta | \mathbf{x}) &= \frac{p_{\mathbf{X}, \theta}(\mathbf{x}, \theta)}{p_{\mathbf{X}}(\mathbf{x})}, \\ &\propto_{\theta} p_{\mathbf{X}, \theta}(\mathbf{x}, \theta) \end{aligned} \quad (11)$$

since $p_{\mathbf{X}}(\mathbf{x})$ is independent of θ . $p_{\mathbf{X}, \theta}(\dots)$ is, of course, the joint density of \mathbf{X} and θ . Again by definition we have

$$\begin{aligned} p_{\mathbf{X}}(\mathbf{x} | \theta) &= \frac{p_{\mathbf{X}, \theta}(\mathbf{x}, \theta)}{p_{\theta}(\theta)}, \\ \implies p_{\mathbf{X}, \theta}(\mathbf{x}, \theta) &= p_{\mathbf{X}}(\mathbf{x} | \theta) \cdot p_{\theta}(\theta). \end{aligned} \quad (12)$$

So combining (11) and (12) we obtain

$$p_{\theta}(\theta | \mathbf{x}) \propto_{\theta} p_{\mathbf{X}, \theta}(\mathbf{x}, \theta) = p_{\mathbf{X}}(\mathbf{x} | \theta) \cdot p_{\theta}(\theta). \quad (13)$$

Usually we take the proportionality of the first and last terms of (13) as the basis of a Bayesian analysis because doing so structures the logical development in the intuitive form:

$$\text{posterior} \propto \text{likelihood} \times \text{prior}.$$

However, here we shall find it convenient to use the proportionality of the first and second terms in (13); that is

$$\text{posterior} \propto \text{joint density along } \mathbf{X} = \mathbf{x}.$$

Until now we have been assuming that the unknown quantities in the physical model have been sufficient to define completely the statistical parameters in the probabilistic description of the experiment, $p_{\mathbf{X}}(\cdot | \theta)$. In other words, we have assumed that the only unknowns upon which the distribution of \mathbf{X} depends are those of the physical model. This is equivalent to assuming that there is no modelling error. Hence we must make the distribution of \mathbf{X} depend on parameters other than those in the physical model. We shall need the following notation.

- θ_2 — the parameters of the physical model; *i.e.* θ_2 represents the unknown quantities which we have previously denoted by an unsubscripted θ .
- θ_1 — the statistical parameters (mean, variance, *etc.*) of the distribution of \mathbf{X} .

With these we may define a structuring of Bayesian analysis known as the three-stage model.

Stage III: Prior knowledge

The scientist's prior knowledge of the parameters in his physical model are represented by the prior density: $p_{\theta_2}(\cdot)$.

Stage II: Modelling error

The scientist's beliefs about the adequacy of his physical model are represented by the probability density: $p_{\theta_1}(\cdot | \theta_2)$. In other words, $p_{\theta_1}(\cdot | \theta_2)$ describes the scientist's relative beliefs in the different possible

values of the statistical parameters of X given the particular values θ_2 of the unknowns in his physical model.

Stage I: Observation error

The observations X have a distribution with parameters θ_1 , *i.e.* the probability density of X is $p_X(\cdot | \theta_1)$.

How these distributions may be defined in practical cases will be illustrated in the next section.

Combining these three densities gives the joint density of X , θ_1 , and θ_2 :

$$p_{X, \theta_1, \theta_2} = p_X(X | \theta_1) \cdot p_{\theta_1}(\theta_1 | \theta_2) \cdot p_{\theta_2}(\theta_2). \quad (14)$$

After the observation $X = x$ has been made, the posterior joint density of θ_1 and θ_2 conditional on x is given by

$$p_{\theta_1, \theta_2}(\theta_1, \theta_2 | x) \propto_{\theta_1, \theta_2} p_{X, \theta_1, \theta_2}(x, \theta_1, \theta_2), \quad (15)$$

i.e. the posterior joint density of θ_1 and θ_2 is proportional to the joint density of X , θ_1 , and θ_2 along $X = x$. Usually the scientist's interest is centred on θ_2 alone, the unknowns in his physical model. In that case he needs the marginal posterior density:

$$p_{\theta_2}(\theta_2 | x) = \int p_{\theta_1, \theta_2}(\theta_1, \theta_2 | x) \cdot d\theta_1 \quad (16)$$

This marginal density is the probabilistic representation of the synthesis of his prior knowledge and the information inherent in the data, due allowance having been made for modelling error.

The Bayesian three-stage model is discussed in detail by Lindley & Smith (1972), Smith (1973), and French (1978). In the next section we shall illustrate its application to the estimation of a reflection's intensity from diffractometer data collected in step-scan mode. We shall avoid giving technical details

and instead concentrate on illustrating how the probability densities in each of the three stages may be developed so as to fairly represent the available information.

3. A profile-fitting method for the analysis of diffractometer intensity data

For a diffractometer operating in step-scan mode, each reflection is recorded by measuring a sequence of N counts as the machine steps across the peak and its local background. Each count C_i is an observation on the true (mean) count λ_i at the i th step.

Thus

$$C_i \sim P_{C_i}(\cdot | \lambda_i), \quad i = 1, 2, \dots, N, \quad (17)$$

where the notation indicates that each C_i is drawn from a distribution with parameter λ_i . The distributions $P_{C_i}(\cdot | \lambda_i)$ are approximately Poisson ('counting statistics') with means λ_i , but are perturbed slightly through instrument instability and, for extremely intense reflections, saturation counting losses. The latter effect does not occur in the weak data sets of protein crystallography, the area of our experience, so we shall ignore it. However, our analysis could be modified to take account of its presence.

Each λ_i is the sum of two elements: a contribution from the reflection's intensity and a contribution from the local background. Thus

$$\lambda_i = J \cdot \pi(x_i) + b(x_i), \quad i = 1, 2, \dots, N, \quad (18)$$

where J is the integrated intensity; $\pi(x)$ is the peak shape function, so $\int \pi(x) \cdot dx = 1$; x_i is the position in the scan of the i th step; and $b(x_i)$ is the background scatter at x_i . The problem is to obtain an estimate of J together with some indication of the precision of this estimate.

The data available for the estimation of J clearly include the measured counts (c_1, c_2, \dots, c_N), but there are other sources of information which are often overlooked. In short, these are: (i) the local behaviour of the background, which may be predicted from the collection geometry; (ii) the properties expected in the peak shape, *e.g.* continuity and, in many cases, unimodality; (iii) the shape of the peaks already analysed, since peak shape tends to vary only slowly through reciprocal space (Diamond, 1969); (iv) the position within the scan of the last measured peak and the reliability of the diffractometer in moving from one reflection to the next; and (v) for a multiple counter diffractometer, the relative position of the peaks within the simultaneously collected scans may be predicted from the collection geometry and, moreover, the peak shapes and backgrounds on these scans will usually be very similar.

Various methods have been proposed for the estimation of J . The majority are summarised and discussed by French (1975), and Oatley & French (1981). None of the methods makes use of all the sources of information listed above; indeed, most base their estimate on the sequence of measured counts alone, ignoring sources (i) to (v) entirely. Furthermore, many lead to positively biased estimates of the intensities and poor, occasionally theoretically incorrect, indications of the precisions of these estimates.

The profile-fitting method which we describe here appears to overcome all the difficulties encountered by other methods. Its development within the framework of the Bayesian three-stage model is entirely natural and straightforward, illustrating the power of the Bayesian approach to organise thought.

Expressions (17) and (18) indicate that the expected value of C_i

$$E(C_i) = \lambda_i = J \cdot \pi(x_i) + b(x_i) \quad (19)$$

for $i = 1, 2, \dots, N$. Our approach is to fit the vector of observed counts with a function of the form

$(J.\pi(x, \alpha) + b(x, \beta))$, where $\pi(x, \alpha)$ and $b(x, \beta)$ are parametric approximations to the true, but unknown peak shape and background functions respectively. Since the peak shape function should integrate to unity, obvious candidates for $\pi(x, \alpha)$ are probability density functions. We have found that Johnson's (1949) suggestion for transforming the normal density curve leads to families of curves which well approximate the peak shapes that arise in protein crystallography. In particular, this choice for $\pi(x, \alpha)$ embodies the information that the underlying peak shape is continuous and unimodal. Fig. 3 illustrates the wide variety of continuous, unimodal curves that can result from this choice of $\pi(x, \alpha)$.

For the backgrounds we have found that a linear approximation is adequate; viz. $b(x, \beta) = \beta_1 + \beta_2 x$.

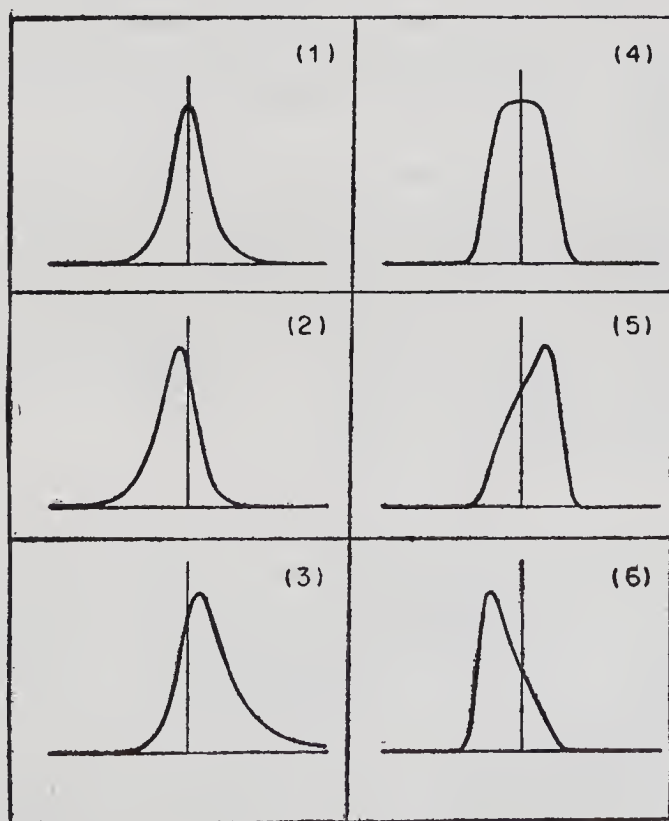


Fig. 3. Examples of peak shapes obtainable for our choice of $\pi(x, \alpha)$.

Usually the background is very nearly constant across the scan, so $\beta_2 \approx 0$.

Other choices of $\pi(x, \alpha)$ and $b(x, \beta)$ may be more appropriate in other branches of crystallography or for other collection geometries. Obviously, if the data indicate that the true peak shape is bimodal, either because the α -doublet is resolved or because the crystal has split, then it is not sensible to use a unimodal choice of $\pi(x, \alpha)$. However, it should be noted that our profile-fitting remains applicable whatever choices are made.

We now set the problem of fitting $(J\pi(x, \alpha) + b(x, \beta))$ to the observed vector of counts into the structure of a Bayesian three-stage model.

Stage III: Prior knowledge

Firstly, let us note the unknown parameters within our model. They are J , the true intensity; α , the parameters of the approximate peak shape function; and β , the parameters of the approximate background function.

Typically we have no prior knowledge of J ; we wish to determine its value from the data alone. We do this through a vague prior distribution. A vague prior distribution is one which states that nothing is known about the relevant quantity other than the information contained in the experimental data. Suppose J can be maximally 10^5 , but is usually of the order of a few hundred. The prior distribution

$$J \sim N(0, 10^{20}), \quad (20)$$

i.e. normal with zero mean and enormous variance, has an effectively constant density over the plausible range of J ; and thus does not differentially weight the possible values of J . Hence the data alone will determine the posterior distribution of J .

The peak parameters α are responsible for determining the position of the peak within the scan, the width of the peak, and properties of the peak shape

such as skewness and kurtosis ('peakedness'). Once several peaks have been fitted, we will have learnt much about these qualities, and hence about α . Suppose that we are setting the prior for the s th reflection, after successfully fitting the profile at the $(s - 1)$ th reflection, adjacent to it in reciprocal space. Suppose further that the posterior distribution for the parameters after the $(s - 1)$ th reflection is

$$\alpha_{s-1} \sim N(\mathbf{m}_{s-1}, W_{s-1}). \quad (21)$$

For our collection geometries there are no predictable changes in α between reflections; but, remembering information sources (iii) and (iv) above, it is known that, first, the peak shape varies only very slowly across reciprocal space and, second, even if there is crystal slippage, the peak position within one scan will be very close to that of the previous reflection. Hence we may use \mathbf{m}_{s-1} as the prior mean for α_s , but should increase the diagonal terms of W_{s-1} slightly to form the prior variance. This slight inflation of the diagonal allows for the unpredictable, but small changes in the peak parameters between reflections.

Finally a prior distribution must be set for β . We shall assume the form: $b(x, \beta) = \beta_1 + \beta_2 x$. In most cases, we use the prior distribution

$$\beta = \begin{pmatrix} \beta_1 \\ \beta_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 10^{20} & 0 \\ 0 & 10^{-10} \end{pmatrix}\right). \quad (22)$$

The variance of 10^{20} corresponds to a vague prior for the background level β_1 , while the variance 10^{-10} forces the background slope to remain very close to zero. If it is believed that the background actually slopes, a more appropriate value than 10^{-10} should be chosen; *e.g.* if β_2 is likely to be in the range -1 to 1 , a prior variance of 1 would be reasonable. Thus through (22) we introduce our prior knowledge of the local behaviour of the background (information source (i)).

Further details of how prior knowledge may be modelled at stage III are discussed by French (1978), and Oatley & French (1981), where ways of treating multiple-counter data are also indicated. We would emphasise that our specific suggestions above are appropriate to our experience within protein crystallography. They may need modifying for other applications; however, the general principles would still hold.

Stage II: Modelling error

This stage describes how well it is expected that the true count λ_i will be modelled by the parametric approximation

$$\nu_i = J \cdot \pi(x_i, \alpha) + b(x_i, \beta). \quad (23)$$

We assume that the approximation is unbiased, *i.e.*

$$E(\lambda_i) = \nu_i. \quad (24)$$

Also it seems reasonable to expect the modelling error to increase with the magnitude of ν_i . In fact, we assume rather more than this: namely, that the modelling error has constant relative variance, *viz.*

$$\text{Var}(\lambda_i | \nu_i) = \sigma_2^2 \cdot \nu_i^2 \quad (25)$$

where σ_2^2 is constant for all steps in the scan. In the solution of the three-stage model these assumptions are modified very slightly. There are computational advantages in working with $\sqrt{\lambda_i}$ and $\sqrt{\nu_i}$ rather than λ_i and ν_i themselves. So we make this transformation and assume that

$$\sqrt{\lambda_i} \sim N(\sqrt{\nu_i}, 0.25 \cdot \sigma_2^2 \cdot \nu_i). \quad (26)$$

In strict probabilistic terms this contradicts (24) and (25), because $E(\sqrt{\lambda_i}) \neq \sqrt{E(\lambda_i)} = \sqrt{\nu_i}$ and because the variance in (26) is not a completely accurate transformation of that given by (25). However, numerically

no great error is introduced (see French (1978), Appendix B).

Further discussion of stage II, in particular of the problem of setting a reasonable value for σ_2^2 , is given by French (1978), and Oatley & French (1981).

Stage I: Observation errors

This stage models the counting and instrument instability errors in the observations. If we temporarily ignore the latter errors, which are small, then we know that $P_{C_i}(\cdot | \lambda_i)$ would be Poisson. Now, if the distribution of C_i is Poisson, the distribution of $\sqrt{C_i}$ may be extremely well approximated by a normal distribution with a constant variance of 0.25 (Box & Taio, 1973, Fig. 1.3.8). Thus we have

$$\sqrt{C_i} \sim N(\sqrt{\lambda_i}, 0.25). \quad (27)$$

The presence of instrument instability error should not disturb this distribution greatly, although it will inflate the variance. McCandlish, Stout & Andrews (1975), amongst others, have argued that machine instability gives rise to errors of approximately constant relative variance in the counts. Letting σ_1^2 be this constant relative variance and assuming that instrument instability is independent of the counting errors, we have

$$\sqrt{C_i} \sim N(\sqrt{\lambda_i}, 0.25 \cdot (1 + \sigma_1^2 \cdot \lambda_i)). \quad (28)$$

This completes our brief description of the three-stage model that underlies our profile-fitting method; full details are given by French (1978), and Oatley & French (1981). There we discuss the very important question of how to set the prior distributions initially, before any reflections have been measured, and also the question of how to set them when the previously fitted reflection was far away in reciprocal space, not adjacent as assumed here.

With all the distributions defined, it is a conceptual-

ly easy, though by no means computationally trivial, task to produce the posterior distribution for J given the observed counts. The mean of this serves as our estimate of the integrated intensity, and the variance indicates the precision.

The method works well in practice; in the past six years it has become the standard method of producing the integrated intensities of diffractometer data in the Laboratory of Molecular Biophysics in Oxford. We present just three examples to illustrate its value here; others may be found in Oatley & French (1978).

In these examples we compare the performance of profile-fitting with that of the ordinate-analysis (Watson *et al.*, 1970) and centroid methods (Tickle, 1975). Both these use the sequence of measured counts to centre a window on the peak. The peak is assumed to lie entirely within the window, and the integrated intensity is taken to be the difference between the total count within the window and the (appropriately scaled) total count outside the window.

Fig. 4 illustrates the profile-fitting of three peaks from a consecutive sequence of eight reflections in some cubic insulin data (Dodson *et al.*, 1978). The ability of the method, guided by its prior knowledge of local behaviour of the background, peak shape, and peak position, to distinguish signal from noise is illustrated in Peak 4. Here ordinate analysis has defined a peak window too near the start of the scan, resulting in overestimation of the intensity. It can also be seen that as a result of crystal movement the peak position has shifted considerably during this sequence, and it is encouraging to note how well this has been tracked by the fitting. This movement has resulted in the last peak, and also the sixth and seventh, being seriously 'clipped'; this would invalidate other methods of step-scan integration, but our profile-fitting method is able to extract valid intensity information from the scan.

Fig. 5 illustrates three reflections taken from some 2-Zn insulin data (Dodson *et al.*, 1979). These

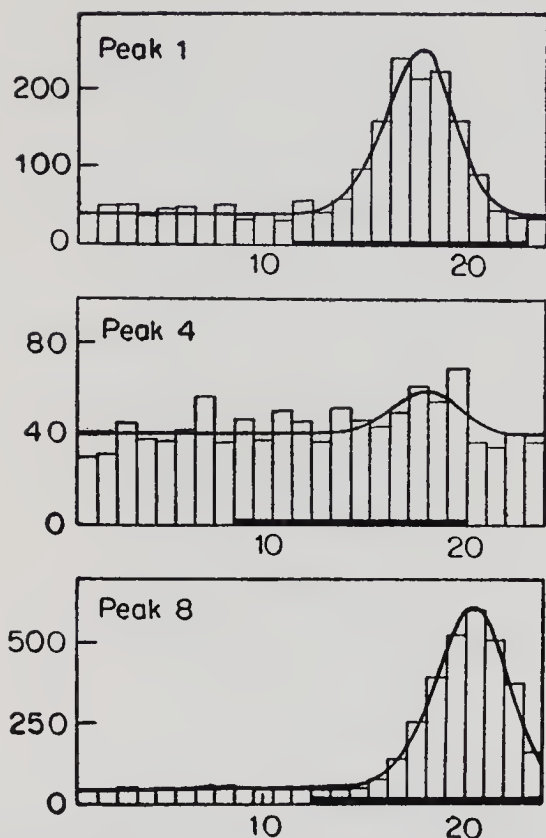


Fig. 4. Profile-fitting of three reflections from cubic insulin data. The window positions determined by ordinate analysis are indicated by the thick line at the base of each diagram.

Peak 1:	Ordinate analysis:	Intensity 938, s.d. 38.
	Profile-fitting:	Intensity 952, s.d. 41.
Peak 4:	Ordinate analysis:	Intensity 132, s.d. 25.
	Profile-fitting:	Intensity 88, s.d. 22.
Peak 8:	Ordinate analysis:	Intensity 2623, s.d. 56.
	Profile-fitting:	Intensity 2723, s.d. 61.

N.B.: Although the calculated standard deviations make ordinate analysis appear the more accurate method in two of the three cases, this is not so. Ordinate analysis standard deviations are based upon a theoretically incorrect formula—it makes no allowance for the random centring of the window—and, moreover, do not include a contribution for instrument instability errors.

reflections lie close, but not adjacent, to each other in reciprocal space. For these centroid and ordinate analysis produced essentially the same results; profile-

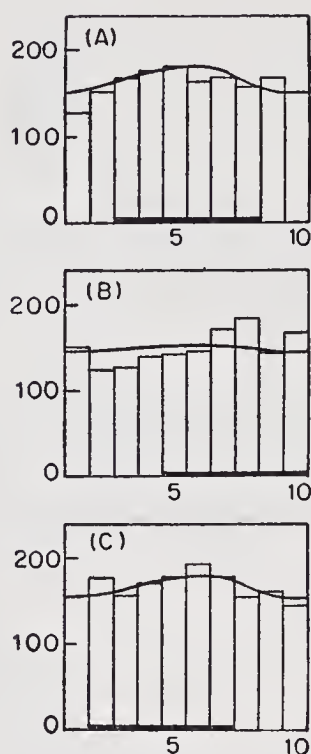


Fig. 5. Profile-fitting of three reflections from 2-Zn insulin data. The window positions determined by ordinate analysis are indicated by the thick line at the base of each diagram.

Peak A: Ordinate analysis:	Intensity 113, s.d. 52.
Profile-fitting:	Intensity 125, s.d. 54.
Peak B: Ordinate analysis:	Intensity 145, s.d. 47.
Profile-fitting:	Intensity 35, s.d. 49.
Peak C: Ordinate analysis:	Intensity 127, s.d. 49.
Profile-fitting:	Intensity 128, s.d. 56.

N.B.: See note to Fig. 4 concerning standard deviations.

fitting agrees with them on the first and third peaks, but produces a much more sensible value for the intensity of the second reflection.

Fig. 6 illustrates the analysis of some data that was collected on a five-counter diffractometer. Ordinate analysis and the centroid method have both been modified to locate the peak window on a combined profile from the five simultaneously collected scans, thus using prior knowledge of the relative positions

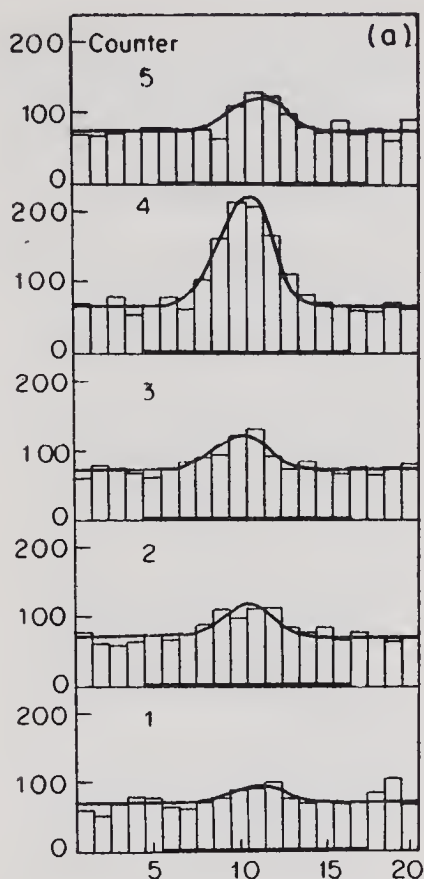


Fig. 6(a). Profile-fitting of reflections from the multiple counter prealbumin data sets. The window positions calculated by the centroid method are indicated by the thick line at the base of each scan. The integrated intensities calculated by the two methods are given below.

Counter:	1	2	3	4	5
Centroid method:	65	229	189	581	175
Profile-fitting:	86	174	182	567	182

of the five peaks (Banner *et al.*, 1977). Because of this, these methods are far more reliable than when used on single-counter diffractometers. Nonetheless, we have found that our profile-fitting can still offer a significant improvement. The three quintuplets shown here are taken from high resolution data for native prealbumin (Oatley, 1976; Blake *et al.*, 1978), and for

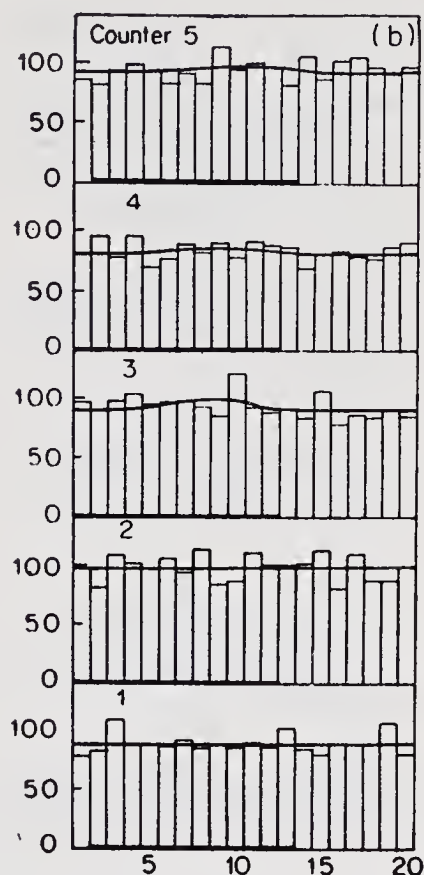


Fig. 6(b). Profile-fitting of reflections from the multiple counter prealbumin data sets. The window positions calculated by the centroid method are indicated by the thick line at the base of each scan. The integrated intensities calculated by the two methods are given below.

Counter:	1	2	3	4	5
Centroid method:	11	-14	93	10	-12
Profile-fitting:	0	4	44	22	24

prealbumin with T_3 and T_4 bound (Oatley & Burrige, 1981). Fig. 6(a) shows a quintuplet where both the centroid method and profile-fitting agree. Fig. 6(b) shows a very weak quintuplet where the centroid method has failed completely to locate the window sensibly. Profile-fitting has not been misled by the generally higher counts in the early part of the scan;

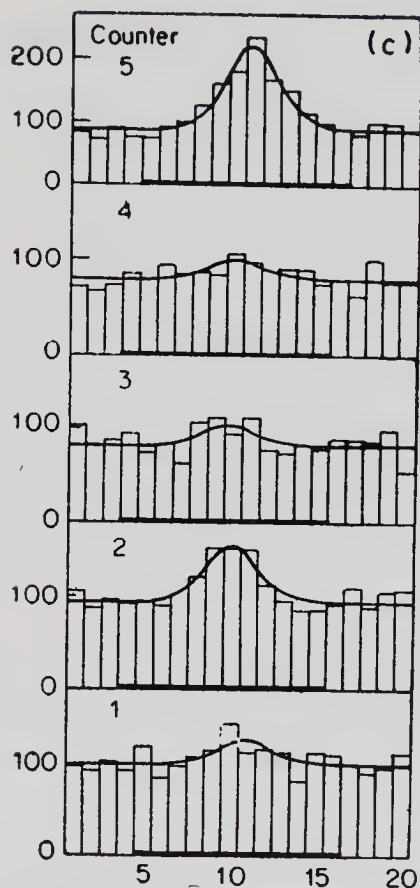


Fig. 6(c). Profile-fitting of reflections from the multiple counter prealbumin data sets. The window positions calculated by the centroid method are indicated by the thick line at the base of each scan. The integrated intensities calculated by the two methods are given below.

Counter:	1	2	3	4	5
Centroid method:	132	144	17	151	564
Profile-fitting:	112	247	95	84	545

to position the peak there would mean moving too far from the positions of previously fitted peaks. Even when the centroid method finds a sensible window position, the random nature of the counts may lead it to produce unsatisfactory integrated intensities. Surely the profile-fitted values are the more sensible for the quintuplet shown in Fig. 6(c).

4. Restrained refinement, hypothesis testing, and Hamilton's *R*-test

It is unusual today to refine the structural parameters of moderate and large molecules against intensity data alone. Restraints (variously known as slack constraints, soft constraints, and pseudo-observations) are introduced into the refinement to encourage, but not force, bond lengths, bond angles, etc. to take sensible values, *i.e.* values similar to those found in previous structural determinations. French (1978) has shown that the Bayesian three-stage model provides a natural framework for discussing such refinements; however, we shall not discuss that here. Instead our concern is with choosing between alternative refined structural models.

Consider a specific example. Suppose that in a protein dimer it is clear that the two molecules are very nearly related by a non-crystallographic two-fold axis. The question is whether they are exactly related so. The approach currently adopted to answer this question is to carry out two restrained refinements: the first with the two molecules constrained to obey the two-fold symmetry relation exactly; the second with the atoms in the two molecules free to move independently, and thus with twice the number of parameters and restraints used in the first refinement. Hamilton's (1965) *R*-test is then applied to see if there is a significant difference between the residuals in the two refinements. Unfortunately the theory of the *R*-test is not applicable to this situation. Neither, for that matter, is the theory of any of the alternatives to the *R*-test that have been proposed (Rogers, 1981; Rothstein, Richardson, & Bell, 1978). However, see Critchley (1980) for a sensible, if *ad-hoc*, approach to using Hamilton's *R*-test in such situations.

The reason why these hypothesis tests are inapplicable is that they are designed to check whether a given hypothesis explains a set of data to within experimental error or whether a more general hypo-

thesis is necessary to explain *the same data*. Now restraints are effectively extra experimental observations on the system under investigation; hence French (1978) termed them pseudo-observations. If one molecular model is refined against a different set of restraints to the other, then they are effectively refined against different data sets; and so frequentist hypothesis testing theory does not apply.

As always the problem is that frequentist statistics provide no framework for handling prior knowledge. The Bayesian framework, on the other hand, does allow for prior knowledge; indeed, it insists that the scientist must always know something about the unknowns in his model, be it only whether they are real or complex. Thus we may expect that a Bayesian hypothesis test may be developed to compare different structural models that have resulted from restrained refinements. We shall not develop such a method in any great detail here; however, we shall indicate how it may be done.

We shall suppose that the scientist has two alternative physical models, M_1 and M_2 , which he wishes to compare. These models are to be taken as functional forms involving unknown quantities, and we consider them before any restrained refinement has taken place. We shall let $P_M(M_i)$ be the scientist's prior belief in model M_i before the refinement ($i=1, 2$). We discuss below the contentious issue of how $P_M(M_i)$ may be set numerically. As we have remarked, there will be unknown quantities within each model; let θ_i be those unknowns within M_i ($i=1, 2$). (Here subscripts on θ 's refer to models, not stages.) Within the context of M_i the scientist will have prior beliefs about θ_i ; let $P_{\theta_i}(\cdot | M_i)$ be the prior density representing these beliefs. In other words, $P_{\theta_i}(\cdot | M_i)$ represents his beliefs if for the time being he assumes that M_i is the true model.

Next he considers the experiment with its observations \mathbf{X} . Let $P_{\mathbf{X}}(\cdot | \theta_i, M_i)$ be the density which

describes his expectation of the observations if he assumes that M_i is the true model and that the unknowns take the values θ_i .

After $\mathbf{I}=\mathbf{x}$ has been observed, Bayes' theorem may be invoked to show that the ratio of posterior probabilities of M_1 and M_2 is

$$\frac{P_M(M_1|\mathbf{x})}{P_M(M_2|\mathbf{x})} = \frac{\int P_X(\mathbf{x}|\theta_1, M_1) \cdot p_{\theta_1}(\theta_1|M_1) \cdot d\theta_1 \cdot P_M(M_1)}{\int P_X(\mathbf{x}|\theta_2, M_2) \cdot p_{\theta_2}(\theta_2|M_2) \cdot d\theta_2 \cdot P_M(M_2)} \quad (29)$$

Noting that $P_M(M_1|\mathbf{x}) + P_M(M_2|\mathbf{x}) = 1$, we can calculate the posterior probabilities of the models. These are natural criteria for choosing between M_1 and M_2 . The more the data and prior knowledge support a model the larger will be its posterior probability and the smaller that of the alternative. Various authors have considered the form of (29) for different distributional assumptions: Jeffreys (1961), Lempers (1971), and Smith & Spiegelhalter (1980) are a useful source of reference. It is possible to combine that Bayesian approach to restrained refinement discussed by French (1978) with the above development, and thus produce an alternative to Hamilton's R -test. However, there is a conceptual problem in (29) that we should admit and discuss.

It is immediately apparent in (29) that the ratio of posterior probabilities depends multiplicatively on the ratio of prior probabilities. These prior probabilities are the subjective evaluations of the scientist. Hence critics of the method may argue that it is not scientific; because for a procedure to be scientific it must surely be objective.

There are a number of points to make here, but first let us widen the discussion. All Bayesian inference is subjective. The posterior distribution always depends to some extent on the prior distribution, which represents the scientist's initial subjective beliefs. Thus the methods presented in the earlier sections of this paper are just as liable to the criticism of subjectivity.

We admit the subjectivity inherent in our methods, but do not see it as a failing.

Categorising very broadly, a Bayesian analysis of a problem may fall into one of three classes. First, there might be total agreement amongst the scientific community about what the prior should be. Although it is dangerous to claim the agreement of others, we would suggest that the majority of crystallographers would concur with the use of Wilson's distributions in (2) and (3) and would thus find the analysis of §1 acceptable and, indeed, scientific. Second, although there may be disagreement over the validity of certain prior beliefs, it may happen that the data are so strong that they dominate the analysis and lead to essentially the same posterior distribution whatever reasonable prior distribution is used. Box & Taio, (1973) give an example of this. Third, there may be disagreement over the validity of certain prior beliefs and the data may be weak. In this case the posterior distribution is sensitive to the choice of prior distribution and there will be a separate Bayesian analysis appropriate to each member of the scientific community. We recommend, therefore, that, when there is no obvious consensus prior distribution, the analysis should be carried out and reported for a range of prior distributions. In that manner it will be possible to see how far the data resolve the initial disagreement and, conversely, how far they leave the controversy open. Thus for the suggested Bayesian hypothesis test the scientist should report the range of $P_M(M_1)$ for which $P_M(M_1 | \mathbf{x}) \geq 0.95$ (say).

It seems to us that the explicit subjectivity of the Bayesian approach is an advantage in that it quickly indicates those aspects of a problem in which the data resolve any disagreement and those aspects where further data must be collected before any agreement can be reached.

We are grateful to many people for their encouragement, advice, criticism, and, not the least, for their

C.S.—4

data. In particular, we should like to thank Professors D. V. Lindley and A. J. C. Wilson, and Doctors S. R. Critchley, R. Diamond, J. S. Rollett, and K. S. Wilson, all who have helped develop the ideas that we have discussed. All members of the laboratory of Molecular Biophysics, past and present, have helped in some way or other, and to them we offer our thanks. S. J. O. is a Mr and Mrs John Jaffé Donation Research Fellow of the Royal Society. Financial assistance was also provided by the Medical Research Council and the Hayward Foundation (S. F.).

References

- BANNER, D. W., EVANS, P. R., MARSH, D. J. & PHILLIPS, D. C. (1977). *J. Appl. Crystallogr.* **10**, 45–51.
- BARNETT, V. (1973). *Comparative Statistical Inference*, New York: John Wiley.
- BLAKE, C. C. F., GEISOW, M. J., OATLEY, S. J., RÉRAT, B. & RÉRAT, C. (1978). *J. Mol. Biol.*, **121**, 339–356.
- BOX, G. E. P. (1980). *J. R. Stat. Soc.* **A143**, 383–430.
- BOX, G. E. P. & TAILO, G. C. (1973). *Bayesian Inference in Statistical Analysis*, Reading, Mass: Addison-Wesley.
- CRITCHLEY, S. R. (1980). D. Phil. Thesis, University of Oxford.
- DEFINETTI, B. (1974). *Theory of Probability*, Vols. 1 and 2, New York: John Wiley.
- DEGROOT, M. H. (1970). *Optimal Statistical Decisions*, New York: McGraw Hill.
- DODSON, E. J., DODSON, G. G., LEWITOVA, A. & SABESAN, M. (1978). *J. Mol. Biol.*, **125**, 387–396.
- DODSON, E. J., DODSON, G. G., HODGKIN, D. C. & REYNOLDS, C. D. (1979). *Can. J. Biochem.*, **57**, 469–479.
- FRENCH, S. (1975). D. Phil. Thesis, University of Oxford.
- FRENCH, S. (1978). *Acta Cryst.* **A34**, 728–738.
- FRENCH, S. & WILSON, K. S. (1978). *Acta Cryst.* **A34**, 517–525.
- HAMILTON, W. C. (1965). *Acta Cryst.* **18**, 502–512.
- HAUPTMAN, H. & KARLE, J. (1953). *Solution of the Phase Problem: I. The Centrosymmetric Crystal*, ACA, Monograph No. 3.
- JEFFREYS, H. (1961). *Theory of Probability*, Oxford University Press.
- JOHNSON, N. L. (1949). *Biometrika* **36**, 149–168.
- LEMPERS, F. B. (1971). *Posterior Probabilities of Alternative Linear Models*, Rotterdam University Press.

- LEWIS, M. L. (1981). Private communication; paper in preparation.
- LINDLEY, D. V. (1965). *Introduction to Probability and Statistics from a Bayesian Viewpoint*, Vols. 1 and 2, Cambridge University Press.
- LINDLEY, D. V. & SMITH, A. F. M. (1972). *J. R. Stat. Soc. B34*, 1-41.
- MCCANDLISH, L. E., STOUT, G. H. & ANDREWS, L. C. (1975). *Acta Cryst. A31*, 245-249.
- MENDES, M. & DE POLIGNAC, C. (1973). *Acta Cryst. A29*, 1-9.
- OATLEY, S. J. (1976). D. Phil. Thesis, University of Oxford.
- OATLEY, S. J. & BURRIDGE, J. M. (1981). Unpublished results.
- OATLEY, S. J. & FRENCH, S. (1981). Paper submitted to *Acta Cryst.*
- ROGERS, D. (1981). *Acta Cryst. A37*, 734-741.
- ROTHSTEIN, S. M., RICHARDSON, M. F. & BELL, W. D. (1978). *Acta Cryst. A34*, 969-974.
- SMITH, A. F. M. (1973). *J. R. Stat. Soc. B35*, 67-75.
- SMITH, A. F. M. & SPIEGELHALTER, D. J. (1980). *J. R. Stat. Soc. B42*, 213-220.
- SRINIVASAN, R. & PARTHASARATHY, S., (1976). *Some Statistical Applications in Crystallography*, Chap. 8. Oxford: Pergamon.
- TICKLE, I. (1975). *Acta Cryst. B31*, 329-331.
- WATSON, H. C., SHOTTON, D. M., COX, J. M. & MUIRHEAD, H. (1970). *Nature (London)* **225**, 806-811.
- WILSON, A. J. C. (1949). *Acta Cryst. 2*, 318-321.
- WILSON, A. J. C. (1981). *Acta Cryst. A37*, 808-810.

Intensity Statistics: Survey, Computer Simulation and the Heavy-Atom Problem

BY URI SHMUELI

*Department of Chemistry, Tel-Aviv University,
Ramat Aviv, 69978 Tel Aviv, Israel*

Abstract

Recent developments in intensity statistics that depend explicitly on the space-group symmetry of the crystal and the atomic composition of the asymmetric unit are illustrated, reviewed and discussed. The need for such generalized statistics is demonstrated by the results of a simple simulation procedure which deals with structure-factor-like summations and their frequency histograms. These simulation exercises confirm that atomic heterogeneity may give rise to serious departures from the ideal (Gaussian) statistics and show that, under such circumstances, different centrosymmetric space groups give rise to entirely different intensity distributions. The mathematical background, required for the derivation of a unified formalism which may account for such real distributions and still be conveniently applicable, is given and some simple but representative derivations are presented. This is followed by a concise but complete summary of all the available relevant expressions. The above formalism is first applied to the simulated distributions, which are satisfactorily accounted for, and is illustrated by its application to a cumulative distribution recalculated from the published structure of a triclinic $C_6N_4O_4Cl_2$ -platinum compound, for which the correct space group ($P\bar{1}$) was rather accurately indicated. Finally, the representations of the generalized probability density functions as Gram-Charlier- and Edgeworth

expansions are examined with regard to their convergence properties. It is concluded that the Gram-Charlier form is to be preferred for practical applications, at least until more terms of the series become available.

1. Introduction

It is well-known that if the asymmetric unit of a crystal contains an outstandingly heavy atom and not too many light ones, a situation frequently encountered in organometallic compounds and other heterogeneous units, the distributions and moments of integrated intensity usually deviate from those predicted by the Wilson (1949) statistics, and resolution of space-group ambiguities with the aid of these ideal statistics often becomes difficult or impossible. The extent of this discrepancy between experimental and ideal statistics is also symmetry-dependent, and statistics which may cope with such situations must therefore take into account both the chemical composition of the asymmetric unit and the space-group symmetry of the crystal.

Probability functions satisfying the above requirements were first given by Karle & Hauptman (1953) and Hauptman & Karle (1953), and were rederived and discussed by other authors (see Srinivasan & Parthasarathy, 1976). However, to the author's knowledge, no applications of these generalized statistics to the heavy-atom problem were published in the 1953–1976 period, presumably because of the apparent complexity of the corresponding equations and the lack of a convenient method whereby space-group symmetry could be accounted for, especially in the case of space groups of symmetries higher than the orthorhombic.

Recent studies of non-ideal intensity statistics (Wilson, 1978; Shmueli, 1979; Shmueli and Kaldor, 1981; Shmueli and Wilson, 1981) confirmed the

earlier results and led to a considerable simplification of the formalism and a generalization of the symmetry treatment. A subsequent extension of the theory, further simplification of the formalism and encouraging applications to distributions based on solved structures, were briefly reported (Shmueli, 1981a; Shmueli, Kaldor & Wilson, 1981) and are described in detail elsewhere (Shmueli, 1982). In the above applications cumulative distributions of normalized structure amplitude, recalculated from several highly heterogeneous asymmetric units (*e.g.*, $C_6 N_4 O_4 Cl_2 Pt, P\bar{1}, Z=2$), were compared with theoretical distributions corresponding to the known compositions and possible space groups, and the known space groups were correctly indicated.

The purpose of the present article is to review the statistical and crystallographic principles underlying these generalized distributions and to discuss the results with particular attention to their potential applicability. Section 2 introduces the heavy-atom problem by means of an easy-to-follow computer simulation of the effects of atomic heterogeneity *and* space-group symmetry on the distribution of structure factors. This, previously unpublished, simulation procedure appears to be of value in preliminary considerations as well as in the assessment of performance of the final expressions. In section 3 generalized distributions, presented as expansions in terms of the ideal distributions (Wilson, 1949) and the associated orthogonal polynomials, are dealt with. A brief survey of the mathematical principles involved (after Cramér, 1951) is followed by a rederivation of the fourth moment of the normalized structure amplitude $|E|$, in terms of the Wilson (1978) statistics of the trigonometric structure factor, and a discussion of explicit relationships of the latter to space-group symmetry operations with particular attention to computational procedures (Shmueli & Kaldor, 1981). This section is concluded with a summary of expressions for probability density functions (p.d.f.), even moments and cumulative

distribution functions (c.d.f.) of $|E|$ which depend on the space-group symmetry and atomic composition and are valid for structures having all the atoms in general positions, with no conspicuous non-crystallographic symmetry and negligible effects of anomalous dispersion.

The section 4 considers the Gram-Charlier and Edgeworth arrangements (Cramér, 1951) of the above expansions, with regard to relationships between the atomic heterogeneity of the asymmetric unit and the rate of convergence of these series. These considerations, of great importance in practical applications, are aided by simulation of distributions corresponding to various heterogeneities and it is tentatively concluded that the Gram-Charlier arrangement should be adopted, at least until more terms of the generalized expansions become available and can be evaluated.

2. A simulation of the heavy-atom problem

The purpose of this section is to introduce and illustrate a method whereby statistical aspects of intensity distribution in a diffraction pattern can be conveniently and reliably simulated. Apart from its probable didactic value, the simulation method to be described affords a means of rapid and extensive numerical tests of the theory with regard to the effects of space-group symmetry and atomic heterogeneity on intensity statistics.

Consider two routine experiments in which intensity data were collected from crystals belonging to space groups $P\bar{1}$ and $Pmmm$. In each case the asymmetric unit contains 24 atoms, all located in general positions, and 3000 non-zero reflexions are available for the structure determination of each crystal. Let us further assume that there is no pseudosymmetry in the structures and that the asymmetric unit of each contains one outstandingly heavy atom and twenty-three

equal lighter ones, the atomic number of the heavy atom being fourteen times that of a light atom in each structure. This ratio of atomic numbers corresponds roughly to one mercury among 23 carbons, which is a rather highly heterogeneous composition.

We wish to simulate the frequency distribution of the structure amplitudes $|F|$, for the above two experiments, and shall first consider the corresponding expressions for the structure factors. These are

$$F_{(1)}(hkl) = 2 \sum_{j=1}^{24} f_j \cos 2\pi (hx_j + ky_j + lz_j) \quad (1)$$

and

$$F_{(2)}(hkl) = 8 \sum_{j=1}^{24} f_j \cos 2\pi hx_j \cos 2\pi ky_j \cos 2\pi lz_j. \quad (2)$$

for $P\bar{1}$ and $Pmmm$ respectively. Since all the atoms occupy general positions, then for a given atomic position $x_j y_j z_j$ and a large set of hkl data, the fractional parts of the products $\mathbf{h} \cdot \mathbf{r}_j = hx_j + ky_j + lz_j$ [eq.(1)] and hx_j, ky_j, lz_j [eq. (2)] are nearly uniformly distributed in the $[0,1]$ range. It is, of course, most important in the present context that the heavy atoms be located in general positions.

It is this uniformity which imparts to the atomic contribution to the structure factor the property of a random variable and, consequently, permits us to regard the structure factor as a sum of random variables. Had we also assumed that all the atoms have the same or similar scattering factors (the 'equal-atom' case), we would be dealing with sums of independent, random variables, all having the same distribution, and the strongest central limit theorem (due to Lindeberg and Lévy, see Cramér, 1951), which predicts a normal distribution of such a sum, would be directly applicable, as is known from the theory and practice of the Wilson (1949) statistics. There are weaker central-limit theorems which predict asymptotic

normality of sums of differently distributed variables, however, the additional conditions they impose reduce the probability that an individual term will have a relatively large contribution to the value of the sum (Cramér, 1951). An increased number of terms clearly achieves this purpose. For example, an outstandingly heavy atom attached to a protein molecule is not expected to give rise to serious departures from the ideal statistics, unlike the situation in small-molecule structures.

The required distributions can be most simply simulated by replacing $hx_j + ky_j + lz_j$, hx_j , ky_j and lz_j by computer-generated pseudo-random numbers p , q , r and s respectively, also uniform in the $[0, 1]$ range, and rewriting (1) and (2) as

$$A_{(1)} = \sum_{j=1}^{24} a_j \cos 2\pi p \quad (3)$$

and

$$A_{(2)} = \sum_{j=1}^{24} a_j \cos 2\pi q \cos 2\pi r \cos 2\pi s \quad (4)$$

respectively, omitting the numerical constants. The composition is simulated by setting $a_1 = 14$ and $a_j = 1$ for $j \neq 1$. However, in order to test the validity of the simulation procedure, distributions for the equal-atom case (all a_j 's equal unity) will also be constructed. The following procedure leads to the required result.

1. Compute the mean $\langle A \rangle$, the absolute deviations from the mean, $\Delta_k = |A_k - \langle A \rangle|$, $k = 1, 3000$ and the variance, $\sigma^2 = \langle \Delta^2 \rangle$, of the distribution of these deviations. Next, construct a histogram of the Δ 's in the $[0, 3\sigma]$ range, collecting their frequency counts in thirty channels, each $\sigma/10$ wide.

The deviations Δ correspond to the magnitudes of the structure factors and the histogram corresponds to their frequency distribution.

2. Compare the histogram with a scaled up ideal p.d.f. (the centric one, in this example)

$$P_c^{(0)}(\Delta) = Kg(\Delta), \quad (5)$$

$$\text{where} \quad g(\Delta) = \left(\frac{2}{\pi\sigma^2}\right)^{1/2} \exp\left(-\frac{\Delta^2}{2\sigma^2}\right) \quad (6)$$

and the scale factor K can be estimated as

$$K = \sum_{i=1}^{30} h(\Delta_i) / \sum_{i=1}^{30} g(\Delta_i), \quad (7)$$

where $h(\Delta_i)$ is the histogram count in the channel centred at Δ_i . Note that the variance of the simulated distribution $h(\Delta)$ is used for the construction of the p.d.f. (6) (or any other p.d.f. to be tested).

The extent of discrepancy between the histogram and the scaled p.d.f. (5) can be expressed by an R factor, defined as

$$R_p^{(c)} = \left\{ \sum_{i=1}^{30} [h(\Delta_i) - P_c^{(0)}(\Delta_i)]^2 / \sum_{i=1}^{30} h^2(\Delta_i) \right\}^{1/2}. \quad (8)$$

3. The 'experimental' cumulative distribution, at the endpoints of the histogram channels, can be computed as

$$N_h(\Delta_k) = \sum_{j=1}^k h(\Delta_j) / 3000 \quad (9)$$

and can be directly compared with the ideal centric and acentric c.d.f.'s

$$N_c^{(0)}(\Delta) = \text{erf}[\Delta/(\sigma\sqrt{2})] \quad (10)$$

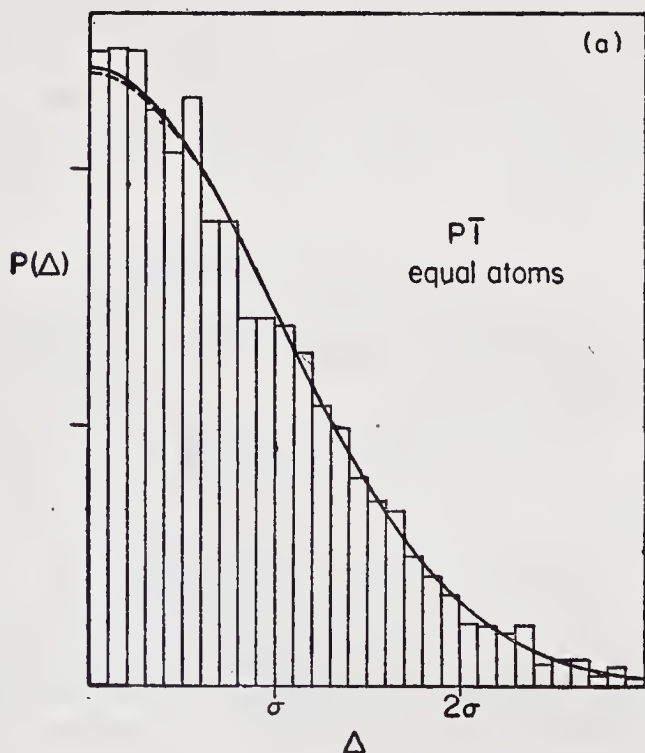
and

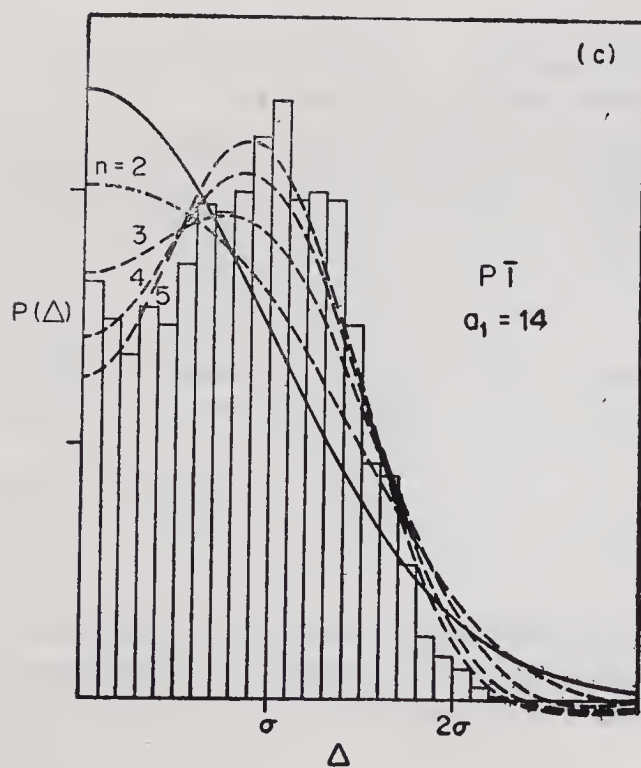
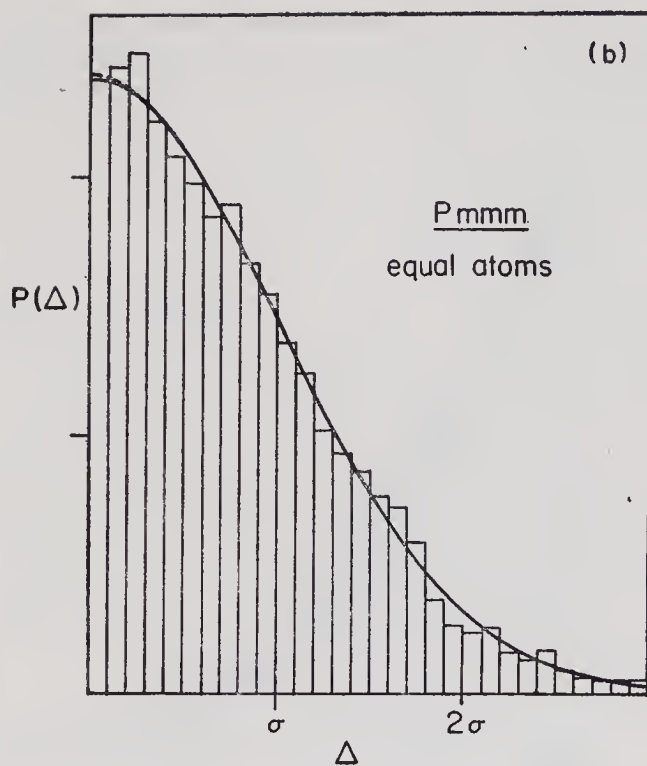
$$N_a^{(0)}(\Delta) = 1 - \exp(-\Delta^2/\sigma^2) \quad (11)$$

based on the Wilson (1949) statistics. Of course, (9) can be compared with other appropriate c.d.f.'s and an R factor analogous to (8) can be evaluated.

The results of our simulation example are contained in Figs. 1 and 2, and some expected and computed statistics, as well as the R factors for the histograms vs. ideal centric p.d.f.'s (5) and the cumulative distributions (9) vs. (10) and (11), are given in Table 1. Figs. 1 and 2 also contain the generalized p.d.f.'s and c.d.f.'s (dashed curves), to be given in the next section, and will be referred to later.

Figs. 1(a) and 1(b) show the distributions obtained for the equal-atom case. The histograms of $P\bar{I}$ [Fig. 1(a)] and $Pm\bar{m}m$ [Fig. 1(b)] agree well with the scaled-up Gaussians, and differences due to different symmetries, or different functional forms of (3) and (4), appear to be unimportant. This was expected since each of (3) and (4) satisfies the assumptions of the Lindeberg-Levy central limit theorem (Cramér, 1951) and these sums should be approximately normally distributed, provided the computer-generated pseudo-random numbers are random enough





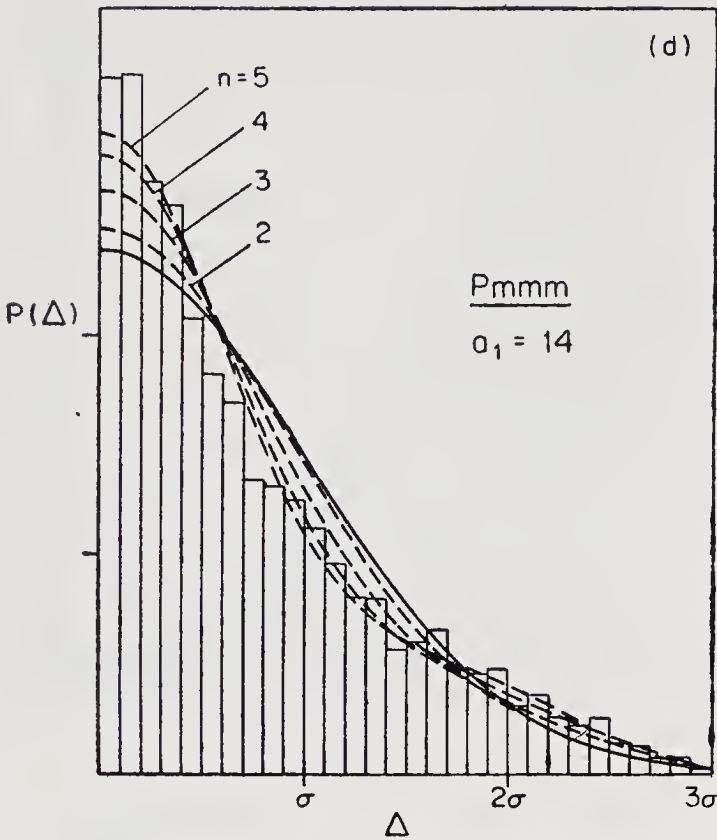


Fig. 1. Comparison of simulated histograms with ideal and non-ideal probability density functions.

The height of each rectangle equals the number of absolute deviations Δ which lie within the corresponding channel. The ideal centric p.d.f.'s (Wilson, 1949) are denoted by solid lines and the non-ideal p.d.f.'s evaluated from a $P(\Delta)$ version of equation (35), are given by the dashed curves. The labels $n = 2$ etc. denote the number of terms of the expansion (35) which were used in the construction of a p.d.f. All the p.d.f.'s are scaled up to the histogram. The variance σ^2 of the histogram is used as the distribution parameter in the ideal and non-ideal p.d.f.'s shown.

(a) $P\bar{1}$ — the equal-atom case, (b) $Pmmm$ — the equal-atom case, (c) $P\bar{1}$ — the one-heavy-atom case, (d) $Pnumm$ — the one-heavy-atom case.

and are sufficiently independent when generated in a contiguous sequence (a single run). The extent of departure from these ideal conditions is rather small,

as can be judged from the values of $\langle A \rangle$ and $\langle A \rangle/\sigma$ which should, of course, be zero (*cf.* Table 1), and the discrepancy between the computed and expected

values of σ . The latter should equal $(\frac{1}{2} \sum_{j=1}^{24} a_j^2)^{1/2}$ and

$(\frac{1}{8} \sum_{j=1}^{24} a_j^2)^{1/2}$ for $P\bar{1}$ and $Pmmm$ respectively. The validity

of this simulation procedure is thus supported by the results shown in Figs. 1(a) and 1(b) and in the 'equal-atom' column in Table 1.

Large departures from normality are seen in the results of the simulations for the heavy-atom case [Figs. 1(c) and 1(d) and Figs. 2(a) and 2(b)]. The departure is much more serious for $P\bar{1}$ ($R_p^{(c)} = 0.394$) than it is for $Pmmm$ ($R_p^{(c)} = 0.212$), and the overall shapes of these distributions are qualitatively

Table 1. *Statistics of simulated histograms and comparison with ideal p.d.f.'s and c.d.f.'s.*

	equal-atom		heavy-atom	
	$P\bar{1}$	$Pmmm$	$P\bar{1}$	$Pmmm$
$\langle A \rangle$	0.0307	-0.0163	0.0432	-0.0778
σ	3.4160	1.7713	10.4454	5.1954
σ_{exp}	3.4641	1.7321	10.4642	5.2321
$\langle A \rangle/\sigma$	0.0090	-0.0092	0.0041	-0.0150
$R_p^{(c)}$	0.063	0.045	0.394	0.212
$R_N^{(\bar{1})}$	0.011	0.005	0.182	0.076
$R_N^{(1)}$	0.186	0.175	0.079	0.241

The results refer to calculations described in text and displayed in Figs. 1 and 2. The means $\langle A \rangle$ are obtained as

$\sum_{i=1}^{3000} A_{(1)}^i/3000$ and $\sum_{i=1}^{3000} A_{(2)}^i/3000$ for $P\bar{1}$ and $Pmmm$ respectively,

and σ is taken as $\langle \Delta^2 \rangle^{1/2}$, as defined in text. For definitions of σ_{exp} (expected)—see text. The R factors are:

$R_p^{(c)}$ —as defined in (8), $R_N^{(\bar{1})}$ —simulated cumulative distribution vs. equation (10) and $R_N^{(1)}$ —simulated cumulative distribution vs. (11)

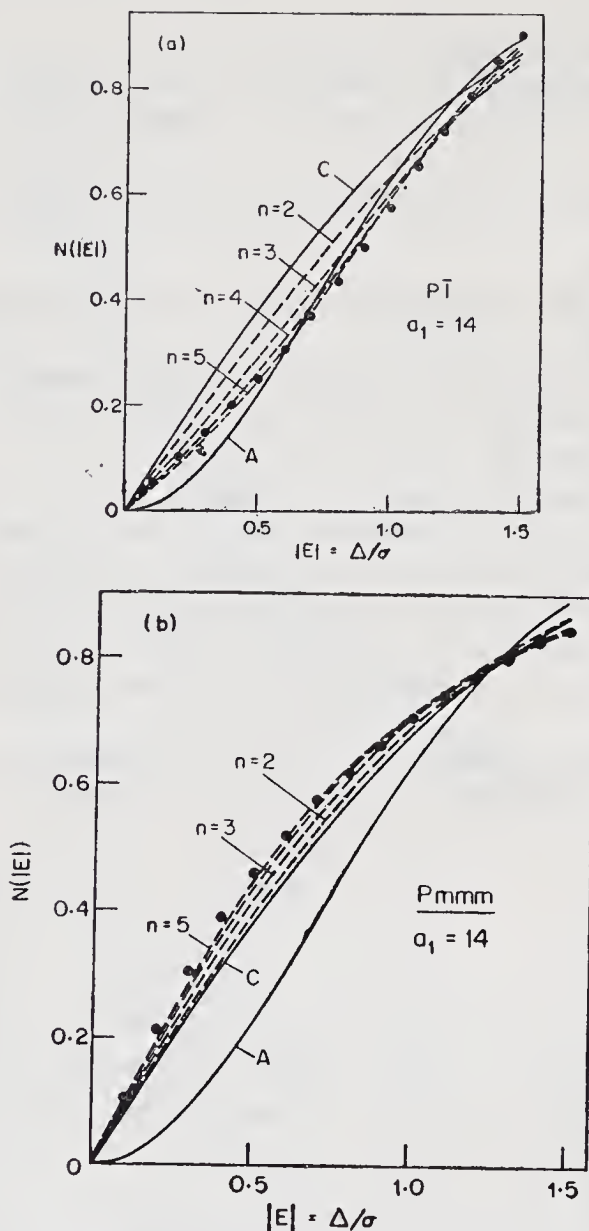


Fig. 2. Comparison of cumulative distributions from the simulated histograms with ideal and non-ideal c.d.f.'s.

The circles denote simulated c.d.f.'s obtained from equation (9), the dashed lines correspond to equation (46) with number of terms indicated in the figure and the solid curves C and A are the ideal centric and acentric c.d.f.'s respectively. The latter are obtained by plotting the r.h.s. of (10) and (11) vs. $|E| = \Delta/\sigma$.

(a) $P\bar{1}$ —the one-heavy-atom case, (b) $Pmmm$ —the one-heavy-atom case.

different. Thus, for one heavy atom in $P\bar{1}$, the frequency (probability) of Δ is at its maximum for $\Delta \approx \sigma$, it is appreciably smaller for Δ near zero and it goes down to zero for $\Delta > 2.3 \sigma$. Since Δ simulates $|F|$, the above description has very much in common with that of the usual distribution of structure amplitudes for a non-centrosymmetric crystal; in fact, the cumulative distribution for the $P\bar{1}$ histogram is seen to be rather close to the ideal acentric c.d.f. [Fig. 2(a), Table 1]. This is consistent with the well-known observation that in presence of a heavy atom in $P\bar{1}$, the distribution 'looks more like an acentric one'.

The high probability for weak intensities (small $|F|$'s or Δ 's), characteristic of the ideal centric distribution, is more accentuated in the simulated histogram for $Pmmm$ in the heavy-atom case, at the expense of frequencies of Δ in the intermediate region [Fig. 1(d)]. This distribution is reminiscent of a hypercentric one and the same impression is given by the corresponding cumulative distribution, which is displaced to the hypercentric side of the ideal centric c.d.f. [Fig. 2(b)].

The important role of space-group symmetry in effects of atomic heterogeneity on intensity statistics has thus been demonstrated for the above two space groups. Such simulations have also been carried out for other centrosymmetric and non-centrosymmetric space groups and a variety of different distributions was obtained in the heavy-atom case. For the equal-atom case, all distributions agree well with the corresponding ideal ones, as in the present example. Several variations of composition were also tried and these can be summarized as follows: (1) an increase of the number of light atoms, for one heavy atom and a fixed $Z_{\text{heavy}}/Z_{\text{light}}$ ratio, and a decrease of this ratio, for a fixed number of atoms in the asymmetric unit, both result in an improved agreement of the simulated statistics with the appropriate ideal one, and (2) the most problematic case appears to be that

C. S.—5

of *one* outstandingly heavy atom among not too many light ones; the presence of two heavy atoms in the asymmetric unit leads to a remarkable decrease in the departure from the ideal statistics and if three or more equal heavy atoms are present, the ideal statistics usually indicates the correct space group quite reliably, even in such an unfavourable space group as $P\bar{1}$ and for the same heterogeneity as the one assumed here (*i.e.* $a_1 = 14$).

Another important application of such simulations is their use in testing theories that are supposed to explain the departures from the ideal statistics in terms of atomic composition and space-group symmetry, *i.e.* the causes of such departures. The theory given in the next section results in truncated series for the probability density functions and a most pertinent practical question, namely, what is the minimum number of terms which are likely to give a correct representation of the non-ideal distribution (or: how many terms must be evaluated?) is examined with the aid of such methods in the last section.

3. Generalized intensity statistics

The heavy-atom problem, illustrated by the above simulations, is an example of a situation which cannot be satisfactorily accounted for in terms of a 'universal' distribution law, in our case the normal distribution. The departures from normality depend, as illustrated above, on the composition of the asymmetric unit as well as on the space-group symmetry of the crystal and hence the required non-ideal or generalized distribution must take these factors into account. Clearly, statistical as well as crystallographic considerations are needed for the derivation of such distributions which will, of course, differ for centrosymmetric and non-centrosymmetric space groups.

A device of mathematical statistics, which often permits an easy derivation of generalized distributions,

without the necessity of proceeding from first principles, is the method of expansions in terms of orthogonal-polynomials (*e.g.*, Cramér, 1951). If the *required* distribution is a generalization of a *known* one and the additional factors which are introduced lead to departures from the known distribution, we may try to expand the required function in terms of the known one, polynomials in our variable and coefficients which depend explicitly on the new factors which require this generalization, *i.e.* problem-dependent coefficients. Such a representation of the required probability density function $P(x)$, which departs from a given p.d.f. $P^{(0)}(x)$, due to factors not accounted for by the latter, is given by

$$P(x) = \sum_k C_k f_k(x) P^{(0)}(x), \quad (12)$$

(Cramér, 1951), where c_k are the problem-dependent coefficients and $f_k(x)$ are polynomials, associated with $P^{(0)}(x)$ by the orthogonality relationship

$$\int_{-\infty}^{\infty} f_k(x) f_l(x) P^{(0)}(x) dx = \begin{cases} 1, & \text{if } k = l, \\ 0, & \text{if } k \neq l. \end{cases} \quad (13)$$

Such polynomials are known as orthogonal with respect to the weight function $P^{(0)}(x)$ (*e.g.*, Abramowitz and Stegun, 1972), and their association with the weight function is unique. Thus, *e.g.*, if $P^{(0)}(x)$ is a Gaussian, the associated f_k 's are Hermite polynomials and if $P^{(0)}(x)$ is of the form $\exp(-x)$, it is associated with Laguerre polynomials (Cramér, 1951; these and other examples are to be found in Abramowitz and Stegun, 1972). The above $P^{(0)}$ functions correspond to the ideal centric and acentric p.d.f.'s and these examples are therefore relevant to our applications.

The choice of orthogonal polynomials for such expansions has several advantages, one of them being related to the problem-dependent coefficients.

Multiplying both sides of (12) by $f_m(x)$ and integrating them, assuming that a term-by-term integration is permissible, we obtain

$$C_m = \int_{-\infty}^{\infty} f_m(x) P(x) dx, \quad (14)$$

where use has been made of the orthogonality relationship (13). The problem-dependent coefficients c_k in (12) are thus expectation values of the corresponding orthogonal polynomials and since the latter are of the form $f_k(x) = \sum_n a_n^{(k)} x^n$, the required coefficients must be

$$C_k = \sum_n a_n^{(k)} \langle x^n \rangle, \quad (15)$$

where $a_n^{(k)}$ are the same coefficients which appear in the polynomials $f_k(x)$, and $\langle x^n \rangle$ are moments of the distribution with the density function $P(x)$, *i.e.* moments of x which contain the *required* problem dependence.

The crystallographic considerations needed for the evaluation of the required moments are illustrated by a derivation of the second moment of the integrated and reduced intensity, $|F|^2$. The derivation follows closely that given by Wilson (1978). The structure factor is given by

$$F(\mathbf{h}) = \sum_j f_j J_j(\mathbf{h}), \quad (16)$$

where f_j is the atomic scattering factor and

$$J_j(\mathbf{h}) = \sum_s \exp [2\pi i \mathbf{h}^T (P_s \mathbf{r}_j + \mathbf{t}_s)], \quad (17)$$

is the trigonometric structure factor of the j th atom, the summation in (16) extends over all the atoms in the asymmetric unit and that in (17) ranges over all

the space-group operations ($P_s | t_s$) which transform the position in which the j th atom is located to its symmetry-equivalent positions. In what follows, only the set of general positions will be considered and effects of dispersion will be neglected. (The latter effects are allowed for in Wilson's original derivation.) The second moment of (corrected) intensity $|F|^2$, or the fourth moment of $|F|$, can thus be written as

$$\langle |F|^4 \rangle = \langle (FF^*)^2 \rangle = \sum_a \sum_b \sum_c \sum_d f_a f_b f_c f_d \langle J_a J_b^* J_c J_d^* \rangle. \quad (18)$$

According to the Wilson (1978) statistics of the trigonometric structure factor, (i) the average $\langle J_a J_a^* \rangle = \langle |J_a|^2 \rangle$, taken over a large set of reflections, equals the multiplicity of the Wyckoff position in which the a th atom is located (here, the order of the point group), which we denote by p_a , (ii) the averages $\langle J_a J_b \rangle$ and $\langle J_a J_b^* \rangle$, with $a \neq b$, vanish for centrosymmetric and non-centrosymmetric space groups, and (iii) the average $\langle J_a J_a \rangle$ vanishes for non-centrosymmetric space groups and equals p_a for the centrosymmetric ones since $J_a = J_a^*$ in the latter case and (iii) reduces to (i).

It follows that the non-vanishing terms in (18) must contain even moments of $|J_a|$ and thus

$$\langle |F|^4 \rangle = L \sum_{a \neq b} \sum f_a^2 f_b^2 p_a p_b + \sum_a f_a^4 q_a, \quad (19)$$

where $q_a = \langle |J_a|^4 \rangle$. Following condition (iii) of the J statistics, it can be readily verified that the multiplicity L of the double summation on the r.h.s. of (19) equals 3 or 2 according as the space group is centrosymmetric or non-centrosymmetric respectively. Making use of the identity

$$\sum_a \sum_{a \neq b} x_a x_b = \left(\sum_a x_a \right)^2 - \sum_a x_a^2 \quad (20)$$

and dropping the subscripts from the moments of $|J|$ (this is possible since all the atoms occupy general positions) we obtain

$$\langle |F|^4 \rangle = L p^2 \left(\sum_a f_a \right)^2 + (q - L p^2) \sum_a f_a^4, \quad (21)$$

(Wilson, 1978).

The fourth moment of the normalized structure amplitude $|E| = |F| / \langle |F|^2 \rangle^{1/2}$ is obtained upon dividing both sides of (21) by

$$\langle |F|^2 \rangle^2 = \left(\sum_a f_a^2 p_a \right)^2 = p^2 \left(\sum_a f_a^2 \right)^2. \quad (22)$$

It follows that

$$\langle |E|^4 \rangle = L + (\gamma_4 - L) Q_4, \quad (23)$$

where

$$\gamma_4 = \langle |J|^4 \rangle / \langle |J|^2 \rangle$$

and

$$Q_4 = \sum_a f_a^4 / \left(\sum_a f_a^2 \right)^2.$$

Higher moments of $|E|$ are derived along similar lines. A unified derivation of $\langle |E|^4 \rangle$, $\langle |E|^6 \rangle$ and $\langle |E|^8 \rangle$ was presented by Shmueli and Wilson (1981) and an extension to $\langle |E|^{10} \rangle$ was carried out by Shmueli (1982). Since for the $2n$ th moment a $2n$ -fold summation analogous to (18), must be considered, the algebraic complexity of the derivation increases very quickly with increasing order of the moment. However, this effort is justified by the fact that each additional moment permits to add a (calculable) term to the generalized p.d.f.'s, based on (12), and thus to overcome possible problems related to the convergence of the latter to the observed (or simulated) p.d.f. (see below).

The space-group dependence of the moments is contained in the even moments of the trigonometric structure factor [cf. eqs. (23) and (17)]. Since the real and imaginary parts of J are listed for all the space groups and all the hkl subsets leading to different

functional forms of J (*International Tables*, 1952) these moments can be evaluated by a straightforward averaging of the resulting trigonometric polynomials and their powers. Such a calculation of $\langle |J|^4 \rangle$ was done by Wilson (1978) for all the space groups but two ($Fd3m$ and $Fd3c$, $p = 192$) and thus made possible the evaluation of the fourth moment of $|E|$. However, a straightforward calculation of moments of $|J|$ higher than the fourth appeared to be too cumbersome and a computer algorithm was constructed which yielded the values of $\langle J^4 \rangle$ and $\langle J^6 \rangle$ for all the centred tetragonal (with $p = 32$) and cubic space groups, starting from the usual trigonometric expressions (Shmueli & Kaldor, 1981). The eighth moment of $|J|$ can also be computed for those space groups with the above algorithm. The remaining space groups were dealt with using an algorithm to be described below, taking $\langle |J|^4 \rangle$ as an example. We have, using equation (17),

$$\langle |J|^4 \rangle = \langle (JJ^*)^2 \rangle = \sum_s \sum_t \sum_u \sum_v \langle \exp [i(\phi_{stuv} + \theta_{stuv})] \rangle, \quad (24)$$

where

$$\phi_{stuv} = 2\pi \mathbf{h}^T (P_s - P_t + P_u - P_v) \mathbf{r} \quad (25)$$

and

$$\theta_{stuv} = 2\pi \mathbf{h}^T (\mathbf{t}_s - \mathbf{t}_t + \mathbf{t}_u - \mathbf{t}_v) \quad (26)$$

(Shmueli and Kaldor, 1981).

As shown by these authors, if \mathbf{r} is a general position then the condition to be fulfilled by a non-vanishing term in (24) is that ϕ_{stuv} be identically zero and the value of such a term is given by $\exp(i\theta_{stuv})$. This can be so only if

$$P_s - P_t + P_u - P_v = 0 \quad (27)$$

and a two-step algorithm follows: (i) find $stuv$ for which (27) holds true and (ii) accumulate the corresponding value of $\exp(i\theta_{stuv})$. The extension of the above to any even order is quite straightforward but

this algorithm is less efficient than the former one when dealing with space groups for which p exceeds 24 (Shmueli & Kaldor, 1981).

Equation (27) and its higher analogs permit an easy evaluation of the lower limits of $\langle |J|^{2n} \rangle$ for all the space groups. Thus, (27) *must* hold true if (i) $s = t = u = v$ (there are p such terms) and (ii) $s = t \neq u = v$ or $s = v \neq t = u$ [$2p(p-1)$ terms]. Equation (24) thus contains at least $p + 2p(p-1) = 2p^2 - p$ non-vanishing terms and the lower limit of $\langle |J|^4 \rangle$ is just this number. Defining

$$\gamma_{2k} = \langle |J|^{2k} \rangle / \langle |J|^2 \rangle \quad (28)$$

and making use of the fact that $\langle |J|^2 \rangle$ always equals p , the order of the point group times the lattice multiplicity, we obtain, from similar considerations, the following lower limits.

$$\gamma_4 \geq 2 - \frac{1}{p}, \quad (29)$$

$$\gamma_6 \geq 6 - \frac{9}{p} + \frac{4}{p^2}, \quad (30)$$

$$\gamma_8 \geq 24 - \frac{72}{p} + \frac{82}{p^2} - \frac{33}{p^3}, \quad (31)$$

$$\gamma_{10} \geq 120 - \frac{600}{p} + \frac{1250}{p^2} - \frac{1225}{p^3} + \frac{456}{p^4}. \quad (32)$$

It is interesting that the above inequalities become equalities for space groups with $p < 4$ as well as for some others. This empirical observation is consistent with a similar remark made by Wilson (1978) in connection with the relation between the value of $\langle |J|^4 \rangle$ and the order of the point group.

It should be noted that the trigonometric structure factors for triclinic, monoclinic and orthorhombic space groups (except $Fdd2$ and $Fddd$) are simple

enough to yield closed expressions for γ_{2k} by direct averaging. *e.g.*, for $P\bar{1}$ we have

$$J = 2 \cos 2\pi (hx + ky + lz) \equiv 2 \cos \alpha \quad (33)$$

and

$$\begin{aligned} \langle |J|^{2n} \rangle &= 2^{2n} \langle \cos^{2n} \alpha \rangle, \\ &= 2^{2n} \frac{(2n-1)!!}{(2n)!!}, \end{aligned}$$

(*e.g.*, Abramowitz and Stegun, 1972). Since $p = \langle |J|^2 \rangle = 2$ for $P\bar{1}$, we have

$$\gamma_{2n} = \frac{(2n-1)!!}{n!}. \quad (34)$$

A list of such expressions for the low-symmetry space groups has been given by Shmueli (1982).

We now present the equations of generalized intensity statistics which account for an arbitrary atomic heterogeneity and any space-group symmetry; it is assumed that (i) all the atoms are located in general positions, (ii) there is no pseudosymmetry or other dependence in the structure, and (iii) effects of dispersion are negligible. The probability density functions, to be given below, were constructed from moments of $|E|$, derived as outlined above (Wilson, 1978; Shmueli & Wilson, 1981; Shmueli, 1982), with the aid of the expansion method described at the beginning of this section.

The probability density functions of $|E|$ are given by

$$P_c(|E|) = P_c^{(0)}(|E|) \left\{ 1 + \sum_{k=2}^n \frac{A_{2k}}{2^k(2k)!} H_{2k} \left(\frac{|E|}{\sqrt{2}} \right) \right\} \quad (35)$$

and

$$P_a(|E|) = P_a^{(0)}(|E|) \left[1 + \sum_{k=2}^n \frac{(-1)^k B_{2k}}{k!} L_k(|E|^2) \right] \quad (36)$$

for centrosymmetric and non-centrosymmetric space groups respectively, where

$$P_c^{(0)}(|E|) = \left(\frac{2}{\pi}\right)^{1/2} \exp\left(-\frac{|E|^2}{2}\right) \text{ and } P_a^{(0)}(|E|) = 2|E| \exp(-|E|^2) \quad (37)$$

are the ideal centric and acentric p.d.f.'s of $|E|$, based on the Wilson (1949) statistics, respectively, the coefficients A_{2k} and B_{2k} depend *via* the even moments of $|E|$ on the composition of the asymmetric unit and space-group symmetry of the crystal and H_{2k} and L_k are Hermite and Laguerre polynomials as defined and tabulated by Abramowitz and Stegun (1972).

The currently available expressions for the coefficients are

$$A_4 \text{ or } B_4 = a_4 Q_4, \quad (38)$$

$$A_6 \text{ or } B_6 = a_6 Q_6, \quad (39)$$

$$A_8 \text{ or } B_8 = a_8 Q_8 + U(a_4^2 Q_4^2 - \gamma_4^2 Q_8), \quad (40)$$

$$A_{10} \text{ or } B_{10} = a_{10} Q_{10} + V\gamma_4^2 Q_{10} + W(a_4 a_6 Q_4 Q_6 - \gamma_4 \gamma_6 Q_{10}) \quad (41)$$

with

$$a_{2k} = \sum_{p=2}^k (-1)^{k-p} (k-p)! \alpha_{kp} \gamma_{2p} + (-1)^{k-1} (k-1)! \alpha_{k0}, \quad (42)$$

where

$$\alpha_{kp} = \binom{k}{p} \frac{(2k-1)!!}{(2p-1)!!} \text{ or } \binom{k}{p} \frac{k!}{p!}, \quad (43)$$

with $(2k-1)!! = (2k)!/(2^k k!)$, according as the space group is centrosymmetric or non-centrosymmetric, and

$$Q_{2k} = \sum_{j=1}^m f_j^{2k} / \left(\sum_{j=1}^m f_j^2 \right)^k \quad (44)$$

where m is the number of atoms in the asymmetric unit and f_j are their scattering factors. The space-group constants γ_{2p} are defined as in (28) and the constants U , V and W appearing in (40) and (41) are given by 35, 3150 and 210 and 18, 900 and 100 for centrosymmetric (A_{2k}) and non-centrosymmetric (B_{2k}) space groups respectively. The first five terms of (35) and (36) can thus be evaluated.

The even moments of $|E|$ are related to the coefficients A_{2k} and B_{2k} , as defined by (38) – (41), by

$$\langle |E|^{2k} \rangle = \alpha_{k0} + \sum_{p=2}^k \alpha_{kp} A_{2p} \text{ (or } B_{2p}), \quad (45)$$

where α_{kp} is defined by (40), and the cumulative distributions of $|E|$ are obtained, by direct integration of (35) and (36) as

$$N_c(|E|) = \operatorname{erf}\left(\frac{|E|}{\sqrt{2}}\right) - \frac{2}{\sqrt{\pi}} \exp\left(-\frac{|E|^2}{2}\right) \sum_{k=2}^n A_{2k} H_{2k-1}\left(\frac{|E|}{\sqrt{2}}\right) / [2^k (2k)!] \quad (46)$$

and

$$N_a(|E|) = 1 - \exp(-|E|^2) + \exp(-|E|^2) \sum_{k=2}^n (-1)^k B_{2k} [L_{k-1}(|E|^2) - L_k(|E|^2)] / k!, \quad (47)$$

for centrosymmetric and non-centrosymmetric space groups respectively (Shmueli, 1982).

4. Discussion

The probability density and cumulative distribution functions, given above, are formally convergent expansions (Shmueli and Wilson, 1981) but they are available for use as truncated series, so far containing at most their first five terms. From a practical point

of view, it is thus important to know whether the available terms satisfactorily represent the actual distributions, under circumstances which may require the use of non-ideal statistics. It is also of interest to examine the rate of convergence, as a function of atomic composition of the asymmetric unit and space-group symmetry of the crystal, as this may indicate how many terms must be included in order to achieve a correct representation.

Before proceeding with the latter question, let us return to the simulation exercise described above and examine the comparisons of equations (35) and (46), as they stand, with the simulated histograms and their cumulative distributions respectively [Figs. 1 and 2]. In each case four curves, corresponding to two-, three-, four- and five-term expansions, are plotted along with the histogram and the ideal distribution.

It is seen from Figs. 1(a) and 1(b) that in the equal-atom case the different truncated expansions nearly coalesce and are remarkably close to the ideal centric p.d.f.'s, for both space groups. This was expected since the A_{2k} coefficients tend to zero as the number of equal atoms tends to infinity, and the small differences between the non-ideal and ideal p.d.f.'s reflect the effects of space-group symmetry in the presence of an asymmetric unit of finite size.

The heavy-atom histogram for $P\bar{1}$, and the corresponding c.d.f., [Figs. 1(c) and 2(a)] are satisfactorily explained by four-term and five-term expansions. The three-term series is the shortest to account semi-quantitatively for the pseudo-acentric hump in the histogram and the first two terms are clearly insufficient for this purpose. Convergence of the $P\bar{1}$ expansions appears to be very slow up to the fourth term but improves at the fifth one. Whether or not the convergence continues to improve with an additional term is not yet clear but this seems to be a worthwhile check, even if six-term expansions will never be used in practical applications.

Comparisons of the various expansions with the somewhat 'hypercentric' $Pmmm$ histogram and its c.d.f. are given in Figs. 1(d) and 2(b). As above, four- and five-term expansions explain the simulated histogram quite well, the three-term expansion is the shortest to account for the excess of very small and very large values of Δ , albeit very approximately, and the two-term expansion appears to be too short for this level of heterogeneity.

The different shapes of the non-ideal p.d.f.'s for $\bar{P}\bar{1}$ and $Pmmm$ are due to the different signs of the symmetry terms a_{2k} , as well as their different magnitudes. Thus, e.g., $a_4 = -1.5$ and 0.375 , $a_6 = 10$ and -5 for $\bar{P}\bar{1}$ and $Pmmm$ respectively and similar relationships hold for higher-symmetry terms. It therefore follows that $Pmmm$ expansions are affected by the heavy atom to a lesser extent than those for $\bar{P}\bar{1}$, in accordance with the histogram and analogous distributions recalculated from solved structures (Shmueli, 1981b). However, the rate of convergence is similar for the two space groups and is therefore dependent mainly on the atomic composition of the asymmetric unit.

Let us examine the composition dependent term Q_{2k} for an asymmetric unit containing l carbons and r equal atoms of type X . We have from (44)

$$Q_{2k} = \frac{l f_c^{2k} + r f_x^{2k}}{(l f_c^2 + r f_x^2)^k} = \frac{l + r \rho^{2k}}{(l + r \rho^2)^k}, \quad (48)$$

where $\rho = f_x/f_c$. For the present considerations we may replace the scattering factors by atomic numbers, i.e., $\rho \simeq Z_x/Z_c$, but this is not recommended or needed in applications to real problems. It follows from (48) that for the equal-atom case ($f_x = f_c$ or $\rho = 1$) the composition term is

$$Q_{2k} = \frac{1}{(l + r)^{k-1}}, \quad (49)$$

while for extreme heterogeneity ($\rho \gg 1$) and not too many light atoms ($r \rho^2 \gg l$), we have

$$Q_{2k} \cong \frac{1}{r^{k-1}}. \quad (50)$$

Let us now rewrite the expansion (35) in a symbolic form, in terms of the composition-dependent quantities. Making use of equations (38)–(41), we can write

$$P_c(|E|) = P_c^{(0)}(|E|) S_{GC}, \quad (51)$$

with

$$S_{GC} = 1 + d_4 Q_4 + d_6 Q_6 + (d'_8 Q_8 + d''_8 Q_4^2) \\ + (d'_{10} Q_{10} + d''_{10} Q_4 Q_6) + \dots,$$

where the multipliers d_{2k} , d'_{2k} and d''_{2k} contain Hermite polynomials of order $2k$, space-group constants and numerical coefficients. In the above series, as in equations (35) and (36), the terms are arranged according to the increasing order of the orthogonal polynomials and these series are thus examples of the Gram-Charlier expansion (Cramér, 1951), abbreviated here as GC. A possible disadvantage of the GC arrangement is that the terms are not necessarily arranged according to a decreasing order of magnitude (Cramér, 1951), and when this is so, another form of the expansion in which terms with the same orders are grouped together, is strongly recommended by Cramér. The latter is known as the Edgeworth arrangement (Cramér, 1951), abbreviated here as ED.

As can be seen from (49) and (50), $Q_6 \approx Q_4^2$ and $Q_8 \approx Q_4 Q_6$, and the expansion (51) can be rewritten for the two extreme compositions as

$$S_{ED} = 1 + \frac{d_4}{\mu} + \frac{d_6 + d''_8}{\mu^2} + \frac{d'_8 + d''_{10}}{\mu^3} \\ + \frac{d'_{10} + \dots}{\mu^4} + \dots \quad (52)$$

where μ stands for $l + r$, the number of atoms in the asymmetric unit in the equal atom case, or r , the

number of heavy atoms present. This is an Edgeworth-type arrangement but the fifth term in (52) is incomplete since there are more terms of the order of μ^4 in the higher, so far unavailable terms of the expansions (35) and (36). When μ is not too small, this ED arrangement is a clearly asymptotic expansion and is likely to converge quickly. *e.g.*, the fast convergence of the correction term in the equal-atom case ($\mu = 24$), to values differing only slightly from unity [*cf.* Figs. 1(a) and 1(b)] is due to the inverse power series character of the expansion. It was also seen in simulations, not described here, that the first four terms of S_{ED} give rise to a somewhat better agreement with the histogram than that given by a five-term GC expansion, provided at least two heavy atoms are present. In fact, $\mu = 2$ in equation (52) already promises some convergence (*cf.* also Shmueli and Wilson, 1981). However, in the important case of *one* very heavy atom among not too many light ones, all the composition terms are of the order of unity or decrease very slowly with increasing order, the ED arrangement loses its *a priori* asymptotic character and the rate of convergence depends mainly on the d_{2k} , d'_{2k} and d''_{2k} coefficients in (51) and (52).

Rather extensive numerical tests and simulations were carried out in order to examine the applicability of the two arrangements to a description of non-ideal probability densities and cumulative distributions. A detailed report of the results is beyond the scope of this paper and only the main conclusions are presented.

1. In the equal-atom case, both ED and GC arrangements are very close to the corresponding ideal p.d.f.'s (Wilson, 1949) and the effects of space-group symmetry are very small.
2. In the case of $C_l X_r$ asymmetric units with $\rho > 1$, the histograms are well accounted for by ED and GC arrangements, provided $r \geq 2$. Occasionally,

the ED arrangement is better than the GC one but the advantage seems to be marginal in the cases studied.

3. In the one-heavy-atom case, all the so far simulated histograms could be explained by three-term, four-term or five-term GC expansions. The performance of ED arrangements is somewhat worse in some space groups and quite unacceptable in others, *e.g.*, in $P\bar{1}$ and $P1$.

Hence, in spite of the preferable mathematical properties of the Edgeworth arrangement, it was rejected in favour of the Gram-Charlier arrangement insofar as applications to the heavy-atom problem in intensity statistics are concerned. The inapplicability of the ED form to $P\bar{1}$ is illustrated in Fig. 3 in which we compare $N(|E|)$ distributions based on (i) ideal centric and acentric p.d.f.'s, (ii) five-term GC arrangements for $P\bar{1}$ and $P1$ and (iii) four-term ED arrangements for these space groups with the cumulative distribution of $|E|$, as recalculated from the solved $P\bar{1}$ structure of $C_6N_4O_4Cl_2Pt$ (Faggiani, Lippert & Lock, 1980; Shmueli, 1982)*. It should be pointed out that the small departure of the c.d.f.'s for $P1$ from the ideal acentric c.d.f. is due to the fact that the assumed asymmetric unit for $P1$ is exactly twice the size of the unit of $P\bar{1}$, and contains *two* platinum atoms rather than one heavy atom only, present in $P\bar{1}$. The above picture may change when several additional terms are included in the expansions (35) and (36).

The equations of generalized intensity statistics, given at the end of the previous section are presented in the Gram-Charlier arrangement and form the basis of the author's computer program which is described elsewhere (Shmueli, 1982) and is being further developed.

*On the other hand, a good agreement of the recalculated $N(|E|)$ with that based on a five-term Gram-Charlier expansion for $P\bar{1}$ is evident.

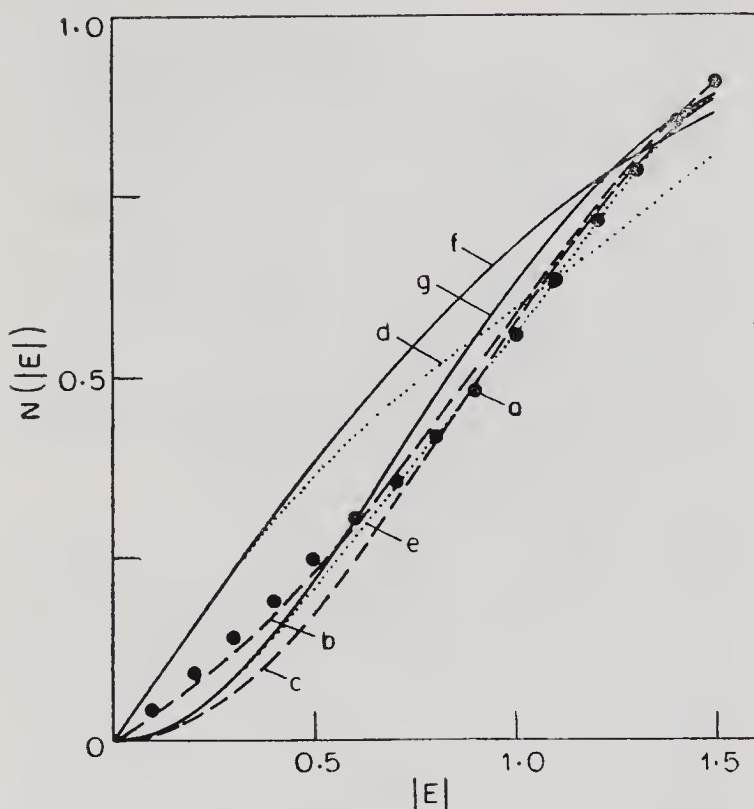


Fig. 3. Cumulative distribution functions for a one-heavy-atom structure.

$N(|E|)$ values, recalculated from the published structure of $C_6N_4O_4Cl_2Pt$ (Faggiani, Lippert & Lock, 1980; Shmueli, 1982), are compared with non-ideal distributions based on the above composition. The correct space group is $P\bar{1}$.

(a) $N(|E|)$ values recalculated from the structure, (b) five-term GC expansion for $P\bar{1}$ (c) five-term GC expansion for $P1$, (d) four-term ED expansion for $P\bar{1}$, (e) four-term ED expansion for $P1$, (f) ideal centric $N(|E|)$ and (g) ideal acentric $N(|E|)$.

The author wishes to thank Professor A. J. C. Wilson for his comments on the simulation approach described above.

The computations were carried out at the Tel-Aviv University Computation Center, with CDC6600 and/or CYBER 172 computers and a NOS/BE operating system.

References

- ABRAMOWITZ, M. & STEGUN, I. A. (1972). *Handbook of Mathematical Functions*. New York: Dover.
- CRAMÉR, H. (1951). *Mathematical Methods of Statistics*. Princeton: University Press.
- FAGGIANI, R., LIPPERT, B. & LOCK, C. J. L. (1980). *Inorg. Chem.* **19**, 295–300.
- HAUPTMAN, H. & KARLE, J. (1953). *Acta Cryst.* **6**, 136–141.
- International Tables for X-ray Crystallography* (1952). Vol. I. Birmingham: Kynoch Press.
- KARLE, J. & HAUPTMAN, H. (1953). *Acta Cryst.* **6**, 131–135.
- SHMUELI, U. (1979). *Acta Cryst.* **A35**, 282–286.
- SHMUELI, U. (1981a). *XIIIth Congress and General Assembly of the IUCr. Contribution 17.X-05*. Ottawa, Canada.
- SHMUELI, U. (1981b). Submitted for publication.
- SHMUELI, U. (1982). *Acta Cryst.* **A38**, 362–371.
- SHMUELI, U. & KALDOR, U. (1981). *Acta Cryst.* **A37**, 76–80.
- SHMUELI, U., KALDOR, U. & WILSON, A. J. C. (1981). *XIIIth Congress and General Assembly of the IUCr. Contribution 17.5-01*. Ottawa, Canada.
- SHMUELI, U. & WILSON, A. J. C. (1981). *Acta Cryst.* **A37**, 342–353.
- SRINIVASAN, R. & PARTHASARATHY, S. (1976). *Some Statistical Applications in X-Ray Crystallography*. Oxford: Pergamon Press.
- WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
- WILSON, A. J. C. (1978). *Acta Cryst.* **A34**, 986–994.

Intensity Statistics: Non-Ideal Distributions in Theory and Practice

BY URI SHMUELI

*Department of Chemistry, Tel-Aviv University,
Ramat Aviv, 69978 Tel-Aviv, Israel*

AND A. J. C. WILSON

*Department of Physics, University of Birmingham,
Birmingham B15 2TT, England*

Abstract

Non-ideal structure-factor statistics are rederived in a general manner and their implementation in practical procedures is indicated and discussed. The derivation employs standard cumulant-moment relationships for a real random variable, as well as some known properties of the cumulants. These properties render the derivation very simple and, at the same time, quite general. The previously assumed vanishing of the odd moments of the trigonometric structure factor is supported by the result of a computation of the third moment of this quantity for a wide range of space groups of high symmetry. Complete five-term expansions for the probability density function of E for centrosymmetric space groups were obtained, without resorting to the assumption of general positions. Previously derived simplifications of these statistics are applied here to the standard even-even cumulant-moment relationships and the results are thereby brought to a very concise functional form, which applies to non-centrosymmetric statistics as well.

Practical aspects of implementation of these statistics in a computer program, including moments of the trigonometric structure factor, angular dependence of the composition-dependent terms and

an efficient arrangement of the computations are discussed and some outstanding problems of theoretical as well as practical interest are briefly indicated.

1. Introduction

Statistical properties of sums of independent random variables can often be represented by ideal distributions which follow from the central-limit theorem (*e.g.*, Cramér, 1951). Such distributions form the basis of the Wilson (1949) statistics and have been extensively used in crystallography. If, however, the relative contributions of these variables to the value of the sum are widely different, *e.g.*, if one of them is outstandingly large, the actual distribution may deviate significantly from the expected ideal one, and such non-ideal distributions may often be approximated by introducing correction terms that depend explicitly on the cause of departure from the ideal situation (Cramér, 1951).

The distribution of diffracted intensity often poses such problems to the crystallographer, and the subject of generalized distributions has been rather extensively investigated (Karle & Hauptman, 1953; Hauptman & Karle, 1953; Bertaut, 1955; Klug, 1958; Foster & Hargreaves, 1963*a*; Srinivasan & Parthasarathy, 1976). However, apart from a few applications of the generalized-moment method (Foster & Hargreaves, 1963*a*, 1963*b*; Goldberg & Shmueli, 1971) no practical use seems to have been made of these non-ideal statistics.

Recent investigations of generalized intensity statistics (Wilson, 1978; Shmueli, 1979; Shmueli & Kaldor, 1981; Shmueli & Wilson, 1981) concentrated mainly on the generalization of their symmetry dependence and on the simplification of the formalism, as these developments appeared to be of major importance in promoting the applicability of such non-ideal statistics. A further simplification, accompanied

by a discussion of some practical aspects and by several encouraging comparisons with distributions which were recalculated from published metallo-organic structures has been given (Shmueli, 1982a) and a summary of these developments has been reported (Shmueli, Kaldor & Wilson, 1981).

This extension of the latter reference aims at (i) standardizing the derivations of the non-ideal statistics, without resorting to the frequently made assumption that all the atoms have to be located in general positions, by applying the conventional cumulant-moment relationships to the centrosymmetric case, and (ii) discussing the implementation of these results in a routine computer program. It is also intended to exploit the simplifications achieved in the former derivations (Shmueli & Wilson, 1981; Shmueli, 1982a) in trying to represent the expressions as concisely as is practicable.

2. Non-ideal intensity statistics

This derivation of generalized moments and distributions for the centrosymmetric case is subject to the assumptions that (i) the atomic contributions to the structure factor may be regarded as independent random variables and (ii) the effects of anomalous dispersion on the intensity distribution are negligible. The contributions in (i) are those of groups of identical atoms related by space-group operations, as given by the atomic trigonometric structure factors, and independence implies that there is no non-crystallographic symmetry within the asymmetric unit. Atoms in *fixed* special positions obviously do not qualify as such contributors and their effect on the distribution must be treated separately.

The structure factor is given by

$$F = \sum_{j=1}^m f_j J_j, \quad (1)$$

where m is the number of atoms in the asymmetric unit, f_j is the atomic scattering factor and J_j is the trigonometric structure factor of the j th atom (Wilson, 1978).

Both F and J_j can here be regarded as real random variables with zero means, and all the odd moments of F vanish. The distribution of F is thus determined by its even moments but can also be represented in terms of the cumulants of F which are related to the moments in a known and often useful way (*e.g.*, Cramér, 1951). Considering general standard relationships between cumulants and moments about the mean (central moments) [*e.g.*, equations (3.43) in Kendall & Stuart, 1969] it is seen that if all the odd moments of a distribution vanish, only even cumulants remain and the first five such cumulants reduce to

$$K_2 = \mu_2, \quad (2)$$

$$K_4 = \mu_4 - 3\mu_2^2, \quad (3)$$

$$K_6 = \mu_6 - 15\mu_4\mu_2 + 30\mu_2^3, \quad (4)$$

$$K_8 = \mu_8 - 28\mu_6\mu_2 + 420\mu_4\mu_2^2 - 630\mu_2^4 - 35\mu_4^2 \quad (5)$$

and

$$K_{10} = \mu_{10} - 45\mu_8\mu_2 + 1260\mu_6\mu_2^2 - 18900\mu_4\mu_2^3 + 22680\mu_2^5 \\ - 210\mu_6\mu_4 + 3150\mu_4^2\mu_2, \quad (6)$$

where K_{2p} and μ_{2p} denote the cumulants and central moments respectively, for a distribution of a real random variable (Kendall & Stuart, 1969).

The cumulants of F , to be denoted by $K_{2p}(F)$, are thus given by the l.h.s. of equations (2)–(6) where the moments μ_{2p} in the r.h.s. of these equations are replaced by the corresponding moments of F . Thus

$$K_2(F) = \langle F^2 \rangle \quad (7)$$

$$K_4(F) = \langle F^4 \rangle - 3\langle F^2 \rangle^2, \quad (8)$$

and so on for the remaining cumulants. These relations will be referred to as the F -version of equations (2)–(6).

According to the additivity of cumulants of

independent random variables (e.g., Cramér, 1951, p. 192), the $2p$ th cumulant of F is given by a sum of $2p$ th cumulants of the atomic contributions $f_J J_J$. Thus

$$K_{2p}(F) = \sum_{j=1}^m K_{2p}(f_J J_J). \quad (9)$$

Another result of the statistical theory tells us that if a random variable, say x , undergoes a linear transformation

$$y = ax + b, \quad (10)$$

where a and b are constants, the r th cumulant of the transformed variable is given by

$$K_r(y) = a^r K_r(x) \quad (11)$$

and is invariant to the shift of the origin (e.g., Kendall & Stuart, 1969, p. 68). Applying this result to equation (9), we have

$$K_{2p}(F) = \sum_{j=1}^m f_J^{2p} K_{2p}(J_J), \quad (12)$$

where the transformation consists of multiplying J_J by the constant f_J .

We can now relate the cumulants of the trigonometric structure factors J_J to their moments, which depend on the space-group symmetry of the crystal (Wilson, 1978; Shmueli & Kaldor, 1981), by constructing a J -version of equations (2)–(6) and the relationships between the cumulants of F and the composition and symmetry of the crystal follow readily from (12). However, before doing so we must note that the even cumulants depend on odd moments as well and the vanishing of the odd moments of J is less obvious than in the case of F . Specifically, terms depending on μ_3^2 , $\mu_5\mu_3$, μ_5^2 and $\mu_7\mu_3$ appear in the complete versions of equations (4), (5) and (6) [cf. Kendall & Stuart, 1969, p. 71]. By analogy with the Wilson (1978) statistics of the trigonometric structure factor, its odd moments are expected to vanish (Shmueli & Wilson, 1981) and were also assumed to

be zero by previous authors (*e.g.*, Karle & Hauptman, 1953; Bertaut, 1955). However, it was stated by Foster & Hargreaves (1963*a*) and recalled by Srinivasan & Parthasarathy (1976) that the mixed odd-even partial moments of the trigonometric structure factor vanish for low symmetries but may be non-zero for symmetries higher than orthorhombic. No demonstration of the truth of this statement was given. This implies that odd moments of J may exist for higher symmetries and such a possibility deserves some consideration, in spite of the fact that odd moments of J are bound to have zero lower limits (*cf.* Shmueli, 1982*b*). In order to test this possibility, the third moment of J was computed by the method of Shmueli & Kaldor (1981) for *all* the trigonal, hexagonal and primitive tetragonal space groups, and was found to vanish for all of them. Although we intend to complete these computations of $\langle J^3 \rangle$ and perform them also for $\langle J^5 \rangle$, we feel justified in using equations (2)–(6) as they stand, in their J -version.

Denoting the first five successive even moments of J by p, q, r, s and t respectively (*cf.* Shmueli & Wilson, 1981) we have, using (12) and the J -version of (2)–(6)

$$K_2(F) = \sum_{j=1}^m f_j^2 p_j, \quad (13)$$

$$K_4(F) = \sum_{j=1}^m f_j^4 (q_j - 3p_j^2), \quad (14)$$

$$K_6(F) = \sum_{j=1}^m f_j^6 (r_j - 15 q_j p_j + 30 p_j^3), \quad (15)$$

$$K_8(F) = \sum_{j=1}^m f_j^8 (s_j - 28 r_j p_j + 420 q_j p_j^2 - 630 p_j^4 - 35 q_j^2) \quad (16)$$

and

$$K_{10}(F) = \sum_{j=1}^m f_j^{10} (t_j - 45 s_j p_j + 1260 r_j p_j^2 - 18900 q_j p_j^3 + 22680 p_j^5 - 210 r_j q_j + 3150 q_j^2 p_j). \quad (17)$$

Comparing (13) and (14) with (7) and (8) respectively we obtain

$$\langle F^2 \rangle = \sum_{j=1}^m f_j^2 p_j, \quad (18)$$

$$\langle F^4 \rangle = \sum_{j=1}^m f_j^4 (q_j - 3 p_j^2) + 3 \langle F^2 \rangle, \quad (19)$$

in agreement with Wilson (1978), and similar comparisons readily lead to general equations for $\langle F^6 \rangle$, $\langle F^8 \rangle$ and $\langle F^{10} \rangle$ that reduce to those given by Shmueli & Wilson (1981) and Shmueli (1982a) for the case of all atoms occupying general positions of a centrosymmetric space group.

The above results can be expressed in terms of the normalized structure factor E by noting that $\langle E^{2k} \rangle = \langle F^{2k} \rangle / \langle F^2 \rangle^k$ and hence

$$K_{2p}(E) = K_{2p}(F) / \langle F^2 \rangle^p. \quad (20)$$

The moments of E can now be readily related to the composition and symmetry of the crystal but the detailed structure of these moments will not be required for the specification of the available terms in the probability density function (p.d.f.) of E . This function is given by

$$P_c(E) = (2\pi)^{-1/2} \exp\left(-\frac{E^2}{2}\right) \left[1 + \sum_{k=2}^N \frac{A_{2k}}{2^k (2k)!} H_{2k}\left(\frac{E}{\sqrt{2}}\right) + \dots \right], \quad (21)$$

(Shmueli & Wilson, 1981; Shmueli, 1982a), where H_{2k} are Hermite polynomials as defined, e.g., by Abramo-

witz and Stegun (1972), and the Gaussian multiplying the 'correction' expansion is based on the Wilson (1949) centric distribution. The moments of E can be related to the expansion coefficients A_{2k} by the integral

$$\langle E^{2k} \rangle = \int_{-\infty}^{\infty} E^{2k} P_c(E) dE \quad (22)$$

which leads to

$$\langle E^{2k} \rangle = (2k-1)!! + \sum_{p=2}^k \binom{k}{p} \frac{(2k-1)!!}{(2p-1)!!} A_{2p}, \quad (23)$$

where $(2k-1)!! = (2k)!/[2^k k!]$ (Shmueli, 1982a). Upon substituting the moments from (23) into the E -version of equations (2)–(6) and expressing the coefficients A_{2p} in terms of the cumulants of E , we obtain

$$\epsilon^2 A_4 = K_4(F), \quad (24)$$

$$\epsilon^3 A_6 = K_6(F), \quad (25)$$

$$\epsilon^4 A_8 = K_8(F) + 35[K_4(F)]^2 \quad (26)$$

and

$$\epsilon^5 A_{10} = K_{10}(F) + 210K_6(F)K_4(F), \quad (27)$$

where $\Sigma = \langle F^2 \rangle$. For the case of all atoms occupying general positions the above results reduce to those obtained by Shmueli & Wilson (1981) and Shmueli (1982a), who used a more lengthy procedure involving a detailed direct treatment of the moments of $|F|$ and $|E|$. The p.d.f. of the magnitude of E is obtained by doubling the normalization constant $(2\pi)^{-1/2}$ and replacing E with $|E|$ throughout equation (21).

The above approach, employing a straightforward application of well-known properties of cumulants and their relationships to central moments, can be extended to the non-centrosymmetric case as well. However, a bivariate distribution must then be

considered and this leads quite naturally to mixed moments and cumulants and hence to a rather complicated algebra, both in the derivation and the simplification of the resulting expressions. On the other hand, the expressions obtained by the use of moments alone (Shmueli & Wilson, 1981; Shmueli, 1982a) have, apart from the p.d.f.'s, identical functional forms for the centrosymmetric and non-centrosymmetric space groups that follow directly from their unified derivations, and are of a comparable complexity to those given above for the centrosymmetric case. Further simplifications were achieved in the direct moment approach (Shmueli, 1982a) and it is interesting to note that they are applicable to the general even-even cumulant-moment relationships given by equations (2)–(6). These can be written as

$$K_{2r} = a_{2r} + a'_{2r}, \quad (28)$$

where

$$a_{2r} = \sum_{p=2}^r (-1)^{r-p}(r-p)! \alpha_{rp} \mu_{2p} \mu_2^{r-p} + (-1)^{r-1}(r-1)! \alpha_{r0} \mu_2^r, \quad (29)$$

$$a'_4=0, a'_6=0, a'_8=-35\mu_4^2, a'_{10}=-210\mu_6\mu_4+3150\mu_4^2\mu_2 \quad (30)$$

and

$$\alpha_{rp} = \binom{r}{p} \frac{(2r-1)!!}{(2p-1)!!} \quad (31)$$

[See also Shmueli, pp. 53–82 above.] Only three numerical constants are left in the first five even-even K - μ relationships, thus enabling a more concise presentation of the cumulants and hence of the whole formalism. *E.g.*, the first five even cumulants of F can now be written as

$$K_{2p}(F) = \sum_{j=1}^m f_j^{2p} (a_{2pj} + a'_{2pj}) \quad (32)$$

where a_{2pj} and a'_{2pj} are given by (29) and (30) respectively, with the moments μ_{2k} replaced by $\langle J_j^{2k} \rangle$ throughout these equations.

We conclude this section with a note on the corresponding expressions for non-centrosymmetric space groups. These non-ideal statistics are based on a p.d.f. of the magnitude of E or F [*cf.* equation (36), Shmueli (1982*b*)], and differ from the centrosymmetric statistics in the definition of a_{rp} , which is given by

$$a_{rp} = \binom{r}{p} \frac{r!}{p!}, \quad (33)$$

and in the values of the numerical constants in (30). The acentric version of the latter is obtained by replacing 35, 210 and 3150 with 18, 100 and 900 respectively (Shmueli 1982*a*, 1982*b*). There are also exact acentric analogues of equations (13)–(17) or their unified form (32), as far as their relationships to the moments of $|F|$ and the expansion coefficients of the acentric p.d.f. are concerned. However, the left-hand sides of the acentric versions of (13)–(17) cannot be simply interpreted as cumulants of $|F|$.

These simple results for the non-ideal statistics of the magnitude of a complex variable, the real and imaginary parts of which are sums of real random variables, contribute to the applicability of generalized intensity statistics and also appear to be of a more general interest.

3. Practical considerations

Applications of the above formalism to the resolution of space-group ambiguities in the case of extreme atomic heterogeneity have been described by Shmueli (1982*a*). It was shown that in cases in which all the atoms, including the outstandingly heavy scatterer, are located in general positions and there is no conspicuous hypersymmetry in the structure, the

appropriately simplified versions of the non-ideal statistics discussed in the previous section lead to a reliable indication of the known space-group symmetry. One of the examples is shown in Fig. 3 of the paper by Shmueli (p. 81 above). The latter paper illustrates the effects of atomic heterogeneity and space-group symmetry on the p.d.f. of the structure amplitude, reviews the mathematical and crystallographic considerations which lead to the construction of the generalized statistics that cope with such effects and discusses in some detail the convergence behaviour of the expansions [(e.g., equation (21)] for the probability density functions of $|E|$. Of the two known arrangements of terms in such expansions, the Edgeworth and the Gram-Charlier forms (Cramér, 1951), the latter was shown to be preferable and all the expansions presented there are given in their Gram-Charlier arrangements (Shmueli, 1982*b*). We now wish to summarize some practical considerations which may be of interest to the reader who wishes to implement these generalized statistics in routine structure determination procedures.

To be specific, let us rewrite explicitly one of the cumulants of E , say $K_8(E)$, using equations (16) and (20) above. We have,

$$K_8(E) = \sum_{j=1}^m f_j^8 (s_j - 28 r_j p_j + 420 q_j p_j^2 - 630 p_j^4 - 35 q_j^2) / \left(\sum_{j=1}^m f_j^2 p_j \right)^4, \quad (34)$$

where p , q , r and s are the second, fourth, sixth and eighth moments of the trigonometric structure factors respectively. Assuming that all the atoms are located in general positions or, in statistical terms, that all the atomic trigonometric structure factors have the same distribution, we can drop the subscripts from

p, q, r and s and, after using (29), equation (34) can be simplified to

$$K_8(E) = (\gamma_8 - a_{43} \gamma_6 + 2a_{42} \gamma_4 - 6a_{40} - 35 \gamma_4^2) Q_8, \quad (35)$$

where

$$\gamma_{2k} = \langle |J|^{2k} \rangle / \langle |J|^2 \rangle^k, \\ Q_{2k} = \sum_{j=1}^m f_j^{2k} / \left(\sum_{j=1}^m f_j^2 \right)^k, \quad (36)$$

and given the γ_{2k} 's and Q_{2k} 's, the cumulants for the centric distribution [cf. equations (24)–(27)] and their acentric analogues can be most readily computed, and the theoretical distributions evaluated.

As for the standardized moment ratios γ_{2k} , which link these statistics to the space-group symmetry, closed expressions are so far available for the triclinic,

Table 1. *Moments of the trigonometric structure factor for triclinic, monoclinic and orthorhombic space groups (except Fdd2 and Fddd).*

The values of γ_{2k} , as defined in (36), are based on closed expressions for these standardized moments, given by Shmueli (1982a). The factors of lattice multiplicity were excluded and must not be used in conventional comparisons. The values of the moments of $|J|$ can be recalculated by noting that $\langle |J|^2 \rangle$ equals the order of the point group. The values of γ_{2k} are valid for structures having all the atoms in general positions and data sets consisting of general reflexions.

Point group(s)	γ_4	γ_6	γ_8	γ_{10}
1	1	1	1	1
$\bar{1}, 2, m$	$1\frac{1}{2}$	$2\frac{1}{2}$	$4\frac{3}{6}$	$7\frac{7}{6}$
$2/m, m\bar{m}2$	$(1\frac{1}{2})^2$	$(2\frac{1}{2})^2$	$(4\frac{3}{6})^2$	$(7\frac{7}{6})^2$
$m\bar{m}m$	$(1\frac{1}{2})^3$	$(2\frac{1}{2})^3$	$(4\frac{3}{6})^3$	$(7\frac{7}{6})^3$
222	$1\frac{3}{4}$	4	$10\frac{3.9}{6.4}$	$30\frac{4.9}{6.4}$

monoclinic and orthorhombic space groups (except *Fdd2* and *Fddd*) (Shmueli, 1982*a*) and we present for convenience the numerical values of γ_4 , γ_6 , γ_8 and γ_{10} , for the above symmetries, in Table 1. These moment ratios are sufficient in order to compute (the available) five-term expansions for the low-symmetry space groups. Values of γ_4 and γ_6 were obtained for all the space groups and all the *hkl* subsets leading to different intensity distributions, by Shmueli & Kaldor (1981) and those of γ_8 will be presented elsewhere (Shmueli & Kaldor, 1982). Hence, statistical tests including the cumulative distribution of $|E|$ as a three-term expansion as well as the fourth and sixth moments of $|E|$, can now be carried out for any symmetry and the availability of moments for five-term expansions should not be long delayed.

A proper computation of the composition-dependent terms should allow for their, albeit not too strong, dependence on $\sin \theta/\lambda$. It was found convenient to place this computation after the data for the Wilson plot have been evaluated since Q_{2k} can then be computed as a weighted average over the shells used in the construction of the plot, the weight being taken as the number of reflexions contained in such a shell.

It therefore appears advisable to incorporate the generalized statistics in a routine which computes normalized structure amplitudes $|E|$ and also evaluates the experimental statistics of this quantity, or to use the output of such a routine (scattering-factor constants, $\sin \theta/\lambda$ ranges and weights and experimental statistics) as an input to a program which deals with generalized intensity statistics and compares the experimental with the possible theoretical distributions.

The procedure adopted by one of us (U.S.) was to modify the locally available program NORMAL (MULTAN 78, Main *et al.*, 1978) so that the required information is output to a file which is re-edited by

adding the required values of γ_{2k} and a few control parameters. This file is then input to the local intensity statistics routine, INSTAT. Of course, the input to such a routine can be reduced to a specification of the space groups for which theoretical statistics have to be computed and the number of the required expansion terms.

There are several extensions of the available theoretical statistics which may be of interest from theoretical as well as experimental standpoints. We are now studying generalized statistics which account for the presence of complex scattering factors (Wilson & Shmueli, 1982). These may take care of significant effects of anomalous dispersion (Wilson, 1980) in highly heterogeneous asymmetric units, and are related to distributions arising from partial centrosymmetry (e.g., Srinivasan & Parthasarathy, 1976). The effect of fixed special positions, already investigated by Karle & Hauptman (1953) and Hauptman & Karle (1953), is of a considerable interest and in need of simplification and generalization, and last, but not least, the effect of variable special positions can be accommodated by the formalism derived in the previous section, with some minor modifications, when the moments of the corresponding trigonometric structure factors become available (*cf.* Wilson, 1978).

References

- ABRAMOWITZ, M. & STEGUN, I. A. (1972). *Handbook of Mathematical Functions*. New York: Dover.
- BERTAUT, E. F. (1955). *Acta Cryst.* **8**, 823–832.
- CRAMÉR, H. (1951). *Mathematical Methods of Statistics*. Princeton: University Press.
- FOSTER, F. & HARGREAVES, A. (1963a). *Acta Cryst.* **16**, 1124–1133.
- FOSTER, F. & HARGREAVES, A. (1963b). *Acta Cryst.* **16**, 1133–1139.
- GOLDBERG, I. & SHMUELI, U. (1971). *Acta Cryst.* **B27** 2164–2173.
- HAUPTMAN, H. & KARLE, J. (1953). *Acta Cryst.* **6**, 136–141.

- KARLE, J. & HAUPTMAN, H. (1953). *Acta Cryst.* **6**, 131–135.
- KENDALL, M. G. & STUART, A. (1969). *The Advanced Theory of Statistics*, Vol. 1, 3rd ed. London: Charles Griffin.
- KLUG, A. (1958). *Acta Cryst.* **8**, 515–543.
- MAIN, P., HULL, S. E., LESSINGER, L., GERMAIN, G., DECLERCQ, J. P. & WOOLFSON, M. M. (1978). *A System of Computer Programs for Solution of Crystal Structures by Direct Methods*. University of York, England.
- SHMUELI, U. (1979). *Acta Cryst.* **A35**, 282–286.
- SHMUELI, U. (1982a). *Acta Cryst.* **A38**, 362–371.
- SHMUELI, U. (1982b). *Crystallographic Statistics* (Indian Academy of Sciences) pp. 53–82.
- SHMUELI, U. & KALDOR, U. (1981). *Acta Cryst.* **A37**, 76–80.
- SHMUELI, U. & KALDOR, U. (1982). In preparation.
- SHMUELI, U., KALDOR, U. & WILSON, A. J. C. (1981). *XIIth Congress and General Assembly of the IUCr*. Contribution 17.5-01. Ottawa, Canada.
- SHMUELI, U. & WILSON, A. J. C. (1981). *Acta Cryst.* **A37**, 342–353.
- SRINIVASAN, R. & PARTHASARATHY, S. (1976). *Some Statistical Applications in X-Ray Crystallography*. Oxford: Pergamon Press.
- WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
- WILSON, A. J. C. (1978). *Acta Cryst.* **A34**, 986–994.
- WILSON, A. J. C. (1980). *Acta Cryst.* **A36**, 945–946.
- WILSON, A. J. C. & SHMUELI, U. (1982). In preparation.

Intensity Statistics and Probability of Validity of Phase Relations

BY G. B. MITRA AND SIKHA GHOSH

*School of Research in X-rays and Structure of Matter,
Department of Physics, Indian Institute of Technology,
Kharagpur 721 302, India*

Abstract

The cumulative intensity distribution function $N(z)$, as well as the probabilities for the sign and phase relationships remaining valid, [P_+ and $P(\phi)$ respectively, as used in direct analytical determination of crystal structures] depend principally on the nature of distribution of structure-factor components.

So far, the Gaussian distribution of structure-amplitude components has been widely used for both these purposes. Recently, however, for $N(z)$ tests, near-Gaussian expressions like Gram-Charlier and Edgeworth series have received extensive attention. Expressions for P_+ and $P(\phi)$ have also been derived on the basis of similar distribution functions. A correlation between these two types of investigation has been attempted in the present work. It is stressed that one cannot empirically or intuitively assume a particular distribution to hold good for a given crystal. The distribution function has to be and can be determined on the basis of the comparison of experimental $N(z)$ values with those calculated on the basis of different distribution functions. Having obtained the best-fitted distribution function, the expression for P_+ or $P(\phi)$ corresponding to this distribution function should be used in programs for direct determination of signs or phases for crystal-structure determination. In course of the present investigation expressions for $N_0(z)$ and $N_1(z)$ as well as for P_+ and $P(\phi)$ respectively based on Edgeworth,

Rayleigh and some other near-Gaussian distributions have been worked out. The case of crystals containing atoms some of which are outstandingly heavy has also received attention and it has been shown that even when the Gaussian type of distribution holds good the distribution is Gaussian with shifted peaks—the peak-shift depending on the difference between the weights of the heavy and light atoms.

1. Introduction

During the early fifties of the present century, important developments in theoretical X-ray crystallography have taken place. Wilson (1949) enunciated the principles of intensity statistics while Sayre (1952) gave the relation connecting the phase of a given reflection with the magnitudes and phases of all other reflections. From this, the phase of one reflection can be expressed in terms of phases of two other reflections with some probability. An expression for this probability has been calculated by Cochran & Woolfson (1955) and by Cochran (1955) for centrosymmetric and non-centrosymmetric crystals respectively on the basis of Gaussian statistics.

Recently for centrosymmetric crystals Giacovazzo (1976) introduced the Gram-Charlier series and obtained an expression for $P_+(s_h s_{h'} s_{h-h'} \approx +1)$, the probability that $s_h s_{h'} s_{h-h'} \simeq 1$. This probability is slightly different from that given by Cochran & Woolfson (1955). The two expressions are as follows:

(a) Cochran & Woolfson (1955) (Gaussian distribution)

$$P_+(s_h s_{h'} s_{h-h'}) \simeq \frac{1}{2} + \frac{1}{2} \tanh \frac{1}{N^{1/2}} |E_h E_{h'} E_{h-h'}| \quad (1)$$

(b) Giacovazzo (1976) (Gram-Charlier Series)

$$P_+(s_h s_{h'} s_{h-h'}) \simeq \frac{1}{2} + \frac{1}{2} \tanh \frac{1}{N^{1/2}} |E_h E_{h'} E_{h-h'}| \\ \times \frac{A}{B}, \quad (2)$$

where A and B are complicated functions of E 's. Here, $E(\mathbf{h}) = 1/(\sum_j f_j^2)^{1/2}$, where $F(\mathbf{h})$ and f_j are the structure factor and the atomic scattering factor for the reflection $(\mathbf{h} = h, k, l)$ respectively. Similarly for non-centrosymmetric crystals expressions for validity of phase relations involving triplets like $\phi_1 + \phi_2 + \phi_3 = 0$ [where $\phi_{\mathbf{h}}$ is given by $F(\mathbf{h}) = |F(\mathbf{h})| \exp(i\phi_{\mathbf{h}})$] and also for quadruplets, quintets, sextuplets, *etc.* have been worked out by Giacovazzo (1976) and Hauptman (1977) based on Gram-Charlier series and Rayleigh distributions respectively. These are different from each other and also from the one derived by Cochran (1955) on the basis of the Gaussian distribution. Again the cumulative probability distribution, $N(z) = \int_0^z p(z) dz$ where $p(z) dz$ is the probability of z lying between z and $z+dz$ and $z = |E|^2$, has been found to be dependent, for centrosymmetric crystals, on whether the distribution is a Gaussian (Wilson, 1949) or an Edgeworth series (Mitra and Belgaumkar, 1973) or a Gram-Charlier series (Shmueli, 1979; Shmueli & Wilson, 1981). Similar results are expected to be valid also for non-centrosymmetric crystals.

2. The Gaussian distribution law—its applications, implications and extensions

The intensity $I(\mathbf{h})$ of a given reflection $\mathbf{h} = (h, k, l)$ is given by $I(\mathbf{h}) = X^2(\mathbf{h}) + Y^2(\mathbf{h})$, (3)

where $X(\mathbf{h}) = \sum_{j=1}^N f_j \cos 2\pi \mathbf{h} \cdot \mathbf{r}_j$,

and $Y(\mathbf{h}) = \sum_{j=1}^N f_j \sin 2\pi \mathbf{h} \cdot \mathbf{r}_j$,

\mathbf{r}_j being the position vector of the j th atom in the unit cell. There are N atoms in the unit cell and j has values ranging between 1 to N . Let us call X and Y the structure-factor components while F is the structure factor. For centrosymmetric crystals $F(\mathbf{h}) = 2 X(\mathbf{h})$

while for non-centrosymmetric crystals equation (3) holds good. By invoking the central limit theorem, Wilson (1949) concluded that $X(\mathbf{h})$ and $Y(\mathbf{h})$ obeyed the Gaussian statistics. Starting from this, Wilson (1949) devised methods of distinguishing between centrosymmetric and non-centrosymmetric unit cells. An improved technique of achieving this end was devised by Howells, Phillips & Rogers (1950). Assuming the Wilson (1949) expression for $p(z)$, they derived expressions for the cumulative probability function $N(z)$ and showed that for a unit cell with no centre of symmetry

$$N_0(z) = 1 - \exp(-z) \quad (4)$$

while for a unit cell with a centre of symmetry

$$N_1(z) = \operatorname{erf} \sqrt{z/2}. \quad (5)$$

Fitting of experimental $N(z)$ values with theoretical curves for $N_0(z)$ and $N_1(z)$ is expected to establish the presence or absence of a centre of symmetry. However, Hargreaves (1955) observed that crystals with heterogeneous atoms—some being outstandingly heavy and the remaining relatively light—show experimental $N(z)$ values not predicted by equation (2) or (3). Hargreaves (1955) correctly attributed the reason to the fact that Wilson (1949) intensity statistics is valid for conglomeration of atoms of same or nearly same atomic number. Theoretical justification of this conclusion lies in the proper enunciation of the central-limit theorem which according to Feller (1969) may be stated as:

Provided that S_1, S_2, \dots, S_n are all independent random variables having the same distribution F and that the average of S_k , $\langle S_k \rangle = 0$ where k is any one of the values of n and that variance of S_k , $\operatorname{var} S_k = 1$, the distribution of the function $S = (S_1 + S_2 + S_3 + \dots + S_n) n^{-1/2}$ tends to be Gaussian in the limit $n \rightarrow \infty$.

The enunciation of the central-limit theorem just quoted is simple, and if the stated conditions are

fulfilled convergence to the limit as n increases is quite rapid. However, the theorem holds even if the variables S_k have different distributions, though convergence to the Gaussian form may be much slower. In such cases a Gram-Charlier or Edgeworth series with several terms may be satisfactory; this point has been discussed in some detail by Shmueli and Wilson (Shmueli, 1979; Shmueli & Wilson, 1981; Shmueli, 1982). The requirement of independence may also be relaxed (French & Wilson, 1978; Wilson, 1981).

Collin (1955) and Sim (1958) considered the case of a crystal containing one heavy atom in the asymmetric unit of the unit cell. The heavy atom was placed at the origin and the unit cell was chosen accordingly. The components of the structure factor $F(\mathbf{h})$ given by

$$F(\mathbf{h}) = f_H + \sum_L f_L \exp(i2\pi \mathbf{h} \cdot \mathbf{r}_L)$$

are now no longer the random variables obeying Gaussian or near-Gaussian statistics, but those of $F(\mathbf{h}) - f_H$ are supposed to be so. The resultant expressions are

$$N_0(z) = (1+r^2) \exp(-r^2) \int_0^z \exp[-(1+r^2)z] I_0[2r(1+r^2)^{1/2}z^{1/2}] dz \quad (6)$$

and

$$N_1(z) = \phi[(1+r^2)^{1/2}z^{1/2}-r] + \phi[(1+r^2)^{1/2}z^{1/2}+r], \quad (7)$$

where

$$r = \frac{f_H}{(\sum_L f_L^2)^{1/2}}, \quad \phi(x) = (2\pi)^{-1/2} \int_0^x \exp(-\frac{1}{2}t^2) dt,$$

$$\text{and } I_0(x) = \frac{1}{\pi} \int_0^\pi \exp(-x \cos \phi) d\phi, \text{ a hyperbolic Bessel}$$

function. All these developments have been made on the assumption of Gaussian distribution for the structure-factor components. Klug (1958) and Bertaut (1955) suggested the use of near-Gaussian distributions like the Gram-Charlier series. Mitra & Belgaumkar

(1973), Shmueli (1979) and Shmueli & Wilson (1981) used Edgeworth and Gram-Charlier series for this purpose.

Along with these developments in intensity statistics, a parallel development was made in the field of evaluation of the probability of phase relations being valid. The Sayre (1952) sign relation $s(\mathbf{h})s(\mathbf{h}')s(\mathbf{h}-\mathbf{h}') \sim 1$ for centrosymmetric crystals, the Cochran (1955) phase relation $\phi_{\mathbf{h}} + \phi_{\mathbf{h}'} + \phi_{\mathbf{h}-\mathbf{h}'} \simeq 0$ and the Hauptman & Karle (1953) tangent relation $\tan \phi_{\mathbf{h}} \approx \tan (\phi_{\mathbf{h}'} + \phi_{\mathbf{h}-\mathbf{h}'})$ can all be derived from the fundamental Sayre (1952) equation

$$G(\mathbf{h}) = \theta_{\mathbf{h}} F(\mathbf{h}) = \Sigma_{\mathbf{h}'} F(\mathbf{h}') F(\mathbf{h}-\mathbf{h}') \quad (8)$$

where

$$G(\mathbf{h}) = \frac{1}{v} \int_v \rho^2(\mathbf{r}) \exp(-i2\pi\mathbf{h}\cdot\mathbf{r}) d\mathbf{v},$$

$$F(\mathbf{h}) = \frac{1}{v} \int_v \rho(\mathbf{r}) \exp(-i2\pi\mathbf{h}\cdot\mathbf{r}) d\mathbf{v},$$

and $\theta_{\mathbf{h}}$ is a ratio connecting the scattering factors of crystals consisting of fictitious atoms with electron distribution $\rho^2(\mathbf{r})$ while the actual crystal consists of real atoms with electron density $\rho(\mathbf{r})$. It is obvious that equation (8) contains the assumption that the scattering factors of all atoms in the crystal are same. Thus, the phase-determining relations are truly valid only for light-atom structures and for crystals containing heavy atoms, these relations will have to be modified. Assuming the Gaussian distribution for structure-factor components, Cochran & Woolfson (1955) derived for the sign relationship for centrosymmetric crystals, the probability of its validity to be the expression in equation (1), while Cochran (1955) derived for non-centrosymmetric crystals the expression for the probability of the phase relations to be valid as

$$P(\phi_{\mathbf{h}}) = \frac{\exp[-2Q \sin^2 \frac{1}{2} (\phi_{\mathbf{h}} - \phi_{\mathbf{h}'} - \phi_{\mathbf{h}-\mathbf{h}'})]}{2\pi \exp(-\phi) I_0(\phi)}, \quad (9)$$

where

$$Q = 2N^{-1/2} |E_{\mathbf{h}} E_{\mathbf{h}'} E_{\mathbf{h}-\mathbf{h}'}|,$$

The above is not and is not meant to be a complete review of all previous work on the subjects mentioned above. The splendid work of Hauptman & Karle (1953) and their later work started with a Rayleigh distribution of structure-factor components for deriving probability of sign and phase relations for triplets, quadruplets, quintets, sextuplets, *etc.* in different neighbourhoods.

Thus it is evident that the probability of phase relations being valid and the cumulative intensity distribution function $N(z)$ both depend on the distribution law of the structure-factor components. Hence it is extremely plausible that a correlation should exist between $N_0(z)$ and $P(\phi)$ and between $N_1(z)$ and $P_+(s_{\mathbf{h}} s_{\mathbf{h}'} s_{\mathbf{h}-\mathbf{h}'})$. The aim of the present work is to investigate this correlation in the case of different distribution laws of the structure factor components. The importance of this correlation in direct determination of crystal structures is too evident to be reemphasized.

3. The Random variables in Gaussian and near Gaussian distributions

It has been seen that for both the sign and phase relationships as well as for the $N(z)$ test, the assumption of atoms with equal weight is implicit. To achieve this for an asymmetric unit with one heavy atom at the origin, Sim (1958) had considered not $F(\mathbf{h})$ but $F(\mathbf{h}) - f_H$ as the random variable. This formalism has been adopted by all subsequent workers in the field. However, this device of removing the heavy atom from the asymmetric unit and considering that the light atoms are randomly distributed is fraught with one glaring mistake. The position of the heavy atom being fixed, this becomes inaccessible to the light atoms

even when the heavy atom is no longer there. The case is now of an outstandingly light atom of weight zero—in the midst of an assembly of light atoms of approximately same weight. This again is a deviation from the Feller (1969) conditions for the validity of the central limit theorem. This can be rectified by filling up every void from which a heavy atom has been removed with a light atom. Let us assume that the positions \mathbf{r}_H occupied by the heavy atoms are known. The structure is then given by

$$F(\mathbf{h}) = \Sigma_H f_H \exp(i 2\pi \mathbf{r}_H \cdot \mathbf{h}) + \Sigma_L f_L \exp(i 2\pi \mathbf{r}_L \cdot \mathbf{h});$$

the f_H 's may have different values while the f_L 's are nearly equal. Σ_H means summation over all heavy atoms while Σ_L means summation over light atoms only and Σ_j represents summation over all atomic positions. If, now, from the known positions \mathbf{r}_H , the heavy atoms are removed and light atoms are placed we have

$$\begin{aligned} F(\mathbf{h}) &= \Sigma_H (f_H - f_L) \exp(i 2\pi \mathbf{r}_H \cdot \mathbf{h}) \\ &= \Sigma_j f_L \exp(2\pi i \mathbf{r} \cdot \mathbf{h}). \end{aligned}$$

In this fictitious structure, all the j atoms are of equal or nearly equal weight and all the 8 atomic positions can be occupied by the atoms of same type with equal probability. Thus, the restricted conditions for the central-limit theorem will be satisfied and $\{F(\mathbf{h}) - \Sigma_H (f_H - f_L) \exp(i 2\pi \mathbf{r}_H \cdot \mathbf{h})\}$ becomes the random variable obeying a Gaussian or near-Gaussian distribution. The distribution in this case as in the case of Sim (1958) is not a pure Gaussian but a shifted Gaussian. In Sim's (1958) case the shift was by f_H . In the present case the peak has been shifted by $\Sigma_H (f_H - f_L) \exp(i 2\pi \mathbf{r}_H \cdot \mathbf{h})$. The standardised structure factor for this fictitious crystal is now

$$\frac{F(\mathbf{h}) - \Sigma_H (f_H - f_L) \exp(i 2\pi \mathbf{r}_H \cdot \mathbf{h})}{(\Sigma_j f_L^2)^{1/2}}, \text{ and}$$

$$z = \frac{|F(\mathbf{h})|^2}{\Sigma_H f_H^2 + \Sigma_L f_L^2}$$

will have to be replaced by $(\rho\sqrt{z} \pm R)^2$
 where $\rho^2 = (r_1^2 + r_2^2)$

$$\text{with } r_1^2 = \frac{\Sigma_H f_H^2}{\Sigma_j f_L^2}, \quad r_2^2 = \frac{\Sigma_L f_L^2}{\Sigma_j f_L^2},$$

$$\text{and } R = \left| \frac{\Sigma_H (f_H - f_L) \exp(i 2\pi \mathbf{r}_H \cdot \mathbf{h})}{(\Sigma_j f_L^2)^{1/2}} \right|.$$

With this formalism, it is easy to see that

$$N_1(z) = \frac{1}{2} [\operatorname{erf}(\rho\sqrt{z} - R)/\sqrt{2} + \operatorname{erf}(\rho\sqrt{z} + R)/\sqrt{2}], \quad (10)$$

and

$$N_0(z) = \exp(-R^2) \int_0^z I_0(2\rho R\sqrt{z}) \exp(-\rho^2 z) dz. \quad (11)$$

For equal atoms, Mitra & Belgaumkar (1973) had derived the following equation assuming the Edgeworth series as the distribution law

$$N_1(z) = \frac{5}{8} \operatorname{erf} \sqrt{\frac{z}{2}} + \frac{3}{2\sqrt{\pi}} \Gamma_{z/2} \left(\frac{1}{2} \right) - \frac{1}{2\sqrt{\pi}} \Gamma_{z/2} \left(\frac{3}{2} \right), \quad (12)$$

where $\Gamma_z(x)$ is the incomplete gamma function.

For crystals with heavy atoms at known positions, it will now be

$$N_1(z) = \frac{5}{16} [\operatorname{erf}(\rho\sqrt{z} - R)/\sqrt{2} + \operatorname{erf}(\rho\sqrt{z} + R)/\sqrt{2}]$$

$$\begin{aligned}
& + \frac{3}{4\sqrt{\pi}} \left[\frac{\Gamma_{(\rho\sqrt{z}-R)^2}(\frac{1}{2})}{2} + \frac{\Gamma_{(\rho\sqrt{z}+R)^2}(\frac{1}{2})}{2} \right] \\
& - \frac{1}{4\sqrt{\pi}} \left[\frac{\Gamma_{(\rho\sqrt{z}-R)^3}(\frac{3}{2})}{2} \right. \\
& \left. + \frac{\Gamma_{(\rho\sqrt{z}+R)^2}(\frac{3}{2})}{2} \right]. \tag{13}
\end{aligned}$$

Figs. 1, 2, 3 and 4 show comparisons of experimental data due to Sim (1958) and Hargreaves (1955) with theoretical curves based on equations due to these authors and the present work. While Fig. 2 shows similar arrangement with both equations (7) and (10) for the rubidium-*o*-nitrobenzoate, Fig. 1 for the potassium-*o*-nitrobenzoate shows a far superior agreement with equation (13), Figs. 3 and 4 also show excellent agreement with equation (10).

From the above it is quite clear that sign and phase relationship for crystals containing heavy atoms will

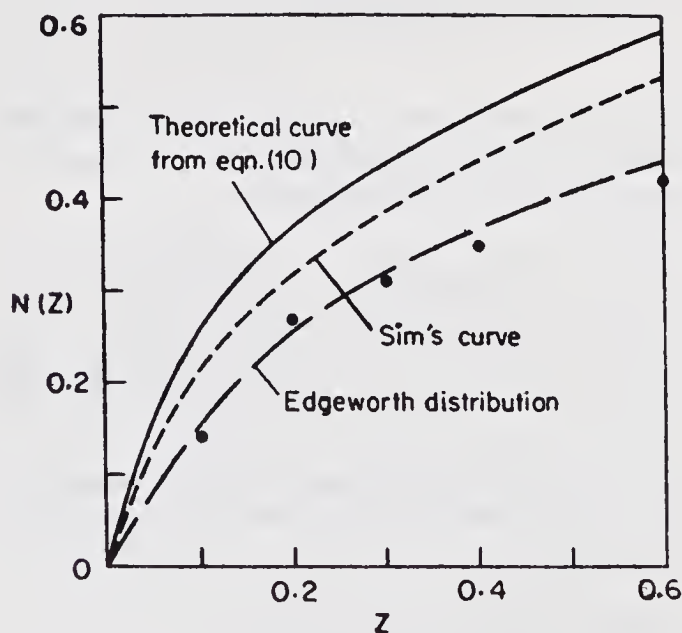


Fig. 1. Comparison of the experimental distribution $N(z)$ for potassium hydrogen di-*o*-nitrobenzoate (marked by circles) with the theoretical distributions and Sim's plot.

have to be modified accordingly. For this case, the sign relationship would be

$$S(\rho E_h - R_h) \approx S(\rho E_{h'} - R_{h'}) S(\rho E_{h-h'} - R_{h-h'}), \quad (14)$$

and the probability that sign of E_h will be the same as that of R_h will be $P = \left\{ \frac{1}{2} + \frac{1}{2} \tanh 2\rho | E_h R_h | \right\}$.

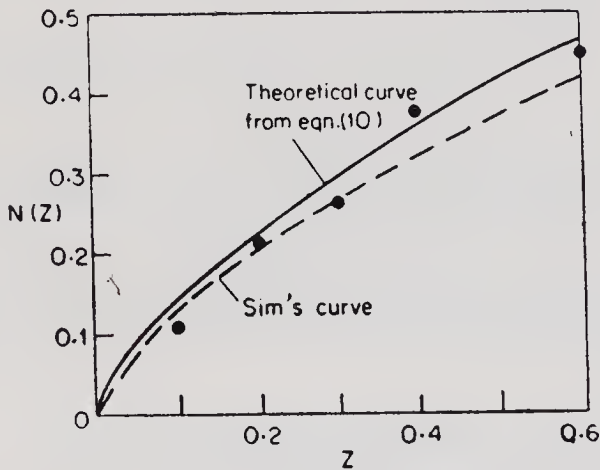


Fig. 2. Comparison of the experimental distribution $N(z)$ for rubidium hydrogen di-*o*-nitrobenzoate (marked by circles) with the theoretical distributions and Sim's plot.

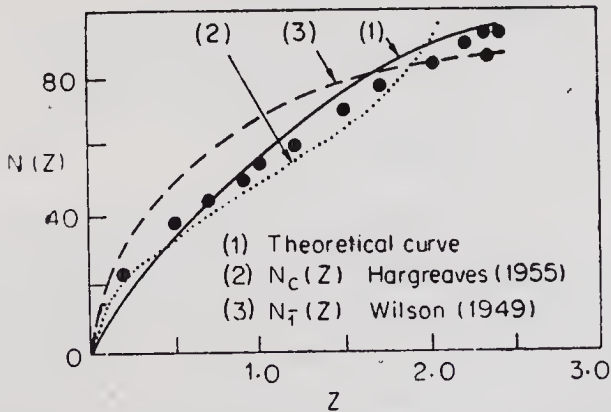


Fig. 3. Comparison of the experimental distribution $N(z)$ for 4-para carbethoxyphenyl 9-stibiafluorene (marked by circles) with the theoretical distributions and Hargreaves plot.

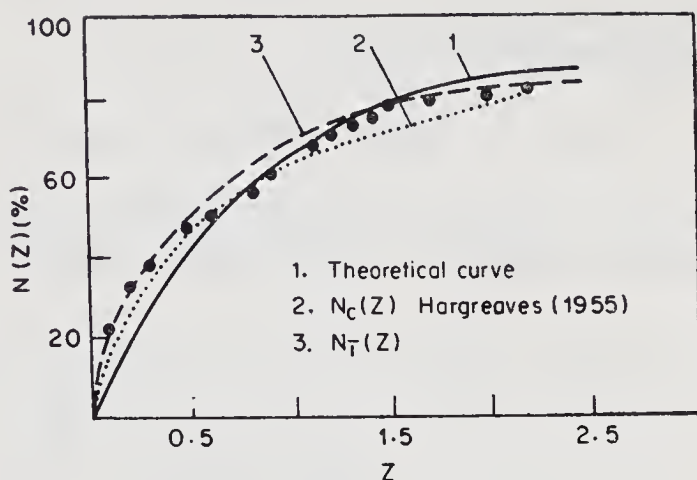


Fig. 4. Comparison of the experimental distribution $N(z)$ for Longifolin hydrobromide (marked by circles) with the theoretical distributions and Hargreaves plot.

Thus the probability that equation (14) will be valid as well as that E_h will be of the same sign as R_h will be

$$P_+ = \left(\frac{1}{2} + \frac{1}{2} \tanh 2\rho |E_h R_h| \right) \left(\frac{1}{2} + \frac{1}{2} \tanh \frac{1}{N^{1/2}} \right. \\ \left. |(\rho E_h - R_h)(\rho E_{h'} - R_{h'}) (\rho E_{h-h'} - R_{h-h'})| \right), \\ \rho^2 = \frac{\Sigma_H f_H^2 \Sigma_L f_L^2}{\Sigma_j f_L^2} \cong \frac{\Sigma_H Z_H^2 + \Sigma_L Z_L^2}{\Sigma_j Z_L^2}; \quad (15)$$

ρ can be considered reasonably independent of h

4. Different types of distributions: their $N(Z)$, P_+ and $P(\phi)$ expressions

The treatment in Section 3 is valid where R_h 's are fully known, *i.e.* the positions of heavy atoms have been unambiguously identified. In many problems this is not so, and then the distribution need not be restricted to Gaussian function, Gram-Charlier series or Edgeworth series. Hauptman and Karle (1953) have used the Rayleigh distribution. Another distribution that suggests itself is the Cauchy distri-

bution, but it is inadmissible, as it predicts an infinite average intensity. The different distributions, their $N_0(z)$, $N_1(z)$, P_+ and $P(\phi)$ expressions are tabulated below.

(a) Gaussian distribution

$$\left. \begin{aligned} N_0(z) &= 1 - \exp(-z) \\ N_1(z) &= \operatorname{erf}(z/2)^{1/2} \end{aligned} \right\} \begin{array}{l} \text{Howells, Phillips \&} \\ \text{Rogers (1950)} \end{array}$$

$$P_+ = \frac{1}{2} + \frac{1}{2} \tanh \frac{1}{N^{1/2}} |E_h E_{h'} E_{h-h'}|$$

Cochran & Woolfson (1955)

$$P(\phi) = \frac{\exp[-2 \sin^2 \frac{1}{2} (\phi_h - \phi_{h'} - \phi_{h-h'})]}{2\pi I_0(x) \exp(-X)}$$

Cochran (1955)

where

$$X = 2N^{-1/2} |E_h E_{h'} E_{h-h'}|$$

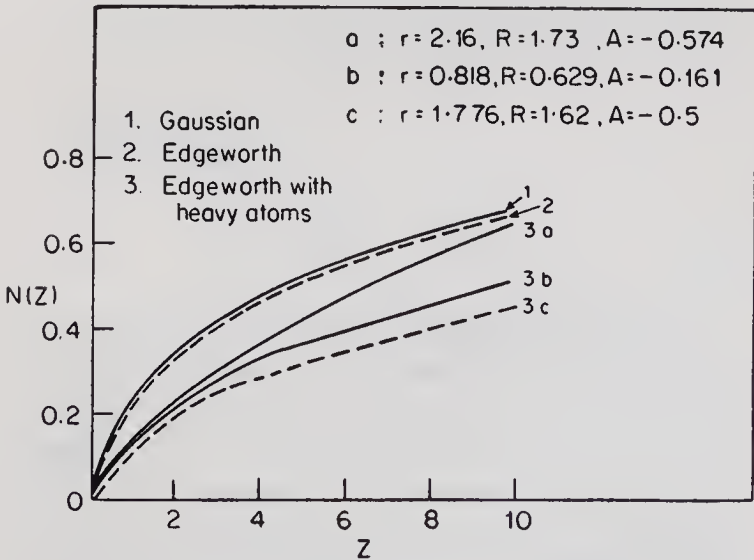


Fig. 5. $N(z)$ plot for centric Gaussian and Edgeworth distributions.

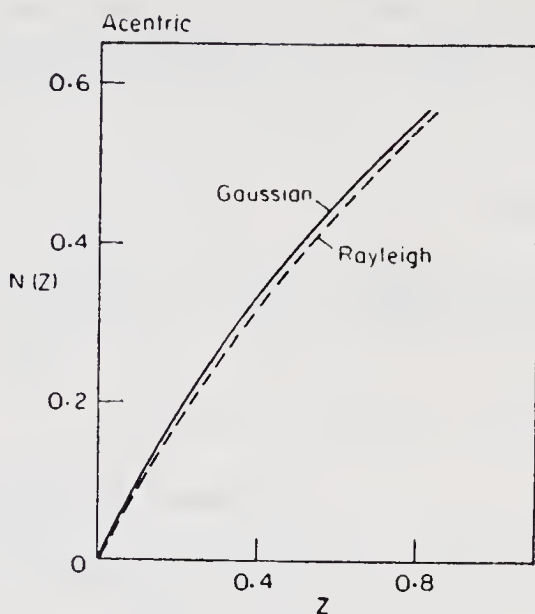


Fig. 6. $N(z)$ plot for acentric Gaussian and Rayleigh distribution.

(b) *Edgeworth distribution* (Mitra & Belgaumkar, 1973)

$$N_1(z) = \frac{5}{8} \operatorname{erf} \sqrt{\frac{z}{2}} + \frac{3}{2\sqrt{\pi}} \Gamma_{z/2} \left(\frac{1}{2} \right) - \frac{1}{2\sqrt{\pi}} \Gamma_{z/2} \left(\frac{3}{2} \right),$$

where $\Gamma_z(x) = \int_0^z t^x \exp(-t) dt$.

Although Mitra & Belgaumkar (1973) had derived expressions for the cumulative distribution function $N(z)$, they did not calculate the probability of validity of P_+ of the sign relationship $s_{\mathbf{h}} s_{\mathbf{h}'} s_{\mathbf{h}-\mathbf{h}'} \approx 1$, holding good when the structure factor components obeyed the Edgeworth-series distribution law. This calculation, which is the prototype of all other calculations to follow, is shown below.

The Edgeworth distribution for normalised structure amplitude is

$$P(E) = \frac{1}{\sqrt{2\pi}} \exp \left[-\frac{(E - \langle E \rangle)^2}{2} \right] \\ \left[\frac{5}{8} + \frac{3}{4} (E - \langle E \rangle)^2 - \frac{1}{8} (E - \langle E \rangle)^4 \right].$$

Evaluating $\langle E \rangle$ in the same way as described by Cochran & Woolfson (1955), after some simplification we obtain

$$P_+(E_h E_{h'}, E_{h-h'}) = \frac{1}{2} + \frac{1}{2} \tanh [2N^{-1/2} E_h E_{h'} E_{h-h'} \\ - \frac{1}{2} \log_e (A+B) + \frac{1}{2} \log_e (A-B)], \quad (16)$$

where

$$A = \frac{5}{8} + \frac{3}{4} (E_h^2 + N^{-1} E_{h'}^2 E_{h-h'}^2) \\ - \frac{1}{8} (E_h^4 - 6 E_h^2 N^{-1} E_{h'}^2 E_{h-h'}^2 + N^{-2} E_h^4 E_{h-h'}^4), \\ B = \frac{3}{2} N^{-1/2} |E_h E_{h'} E_{h-h'}| - \frac{1}{2} N^{-1/2} |E_h^3 E_{h'} E_{h-h'}| \\ - \frac{1}{2} N^{-3/2} |E_h E_{h'}^3 E_{h-h'}^3|$$

(c) *Rayleigh distribution* (Hauptman & Karle 1953)

$$P(R) dR = \frac{2R}{\sigma^2} \exp \left[-\frac{R^2}{\sigma^2} \right] \\ \left\{ 1 - \frac{\sigma_4}{2\sigma_2^2} \left(1 - \frac{2R^2}{\sigma^2} + \frac{R^4}{2\sigma_2^2} \right) \right. \\ \left. - \frac{2\sigma_6}{3\sigma_2^3} \left(1 - \frac{3R^2}{\sigma^2} + \frac{3R^4}{2\sigma_2^2} - \frac{R^6}{6\sigma_2^3} \right) \dots \right\} dR$$

where $R = |F(hkl)|$

Transforming to z and integrating $\int_0^z p(z) dz$, we obtain

$$N_0(z) = 1 - e^{-z} - Az e^{-z} - Bz^2 e^{-z} - Cz^3 e^{-z}, \quad (17)$$

where

$$A = \frac{\sigma_4}{2\sigma_2^2} + \frac{2\sigma_6}{\sigma_2^3}, \quad B = -\frac{\sigma_4}{4\sigma_2^2} - \frac{2\sigma_6}{3\sigma_2^3}, \quad C = \frac{\sigma_6}{9\sigma_2^3},$$

$$\sigma_n = \sum_{j=i}^N f_i^n$$

5. Concluding comments

The distributions discussed are not exhaustive. Many more distributions are still to be explored. The above calculations have been confined to the case of $P1$ and $\bar{P}1$ space groups only. The specialised expressions for different space groups are yet to be calculated.

Again it should be remembered that atomic arrangements in an unit cell are not of the nature of random walk. It is a Markov chain problem. Only one atom in a unit cell may be placed quite randomly. When we bring in the next atom it must be placed on the surface of a sphere equal to the interatomic bond length (*cf.* Wilson, 1981, Patterson interpretation). As we bring in more atoms, position of the n th atom depends on that of $(n-1)$ th atom, whose position depends on that of $(n-2)$ th atom, and so on. $P(n)$ the probability of the n th atom occupying the n th position is then the Markov chain of the conditional probabilities $P(n | n-1 | n-2 | \dots | 2 | 1)$. Our attention and efforts are to be concentrated on this problem to achieve the coveted aim of crystal structure determination by direct analytical methods.

References

- BERTAUT, E. F. (1955). *Acta Cryst.* 8, 823-832.
 COCHRAN, W. (1955). *Acta Cryst.* 8, 473-478.
 COCHRAN, W. & WOOLFSON, M. M. (1955). *Acta Cryst.* 8, 1-12.
 COLLIN, R. L. (1955). *Acta Cryst.* 8, 499-502.

- FELLER, W. (1969). *Introduction to Probability Theory and Its Applications*. New York: Wiley.
- FRENCH, S. & WILSON, K. (1978). *Acta Cryst.* A34, 517-525.
- GIACOVAZZO, C. (1976). *Acta Cryst.* A32, 967-976.
- HARGREAVES, A. (1955). *Acta Cryst.* 8, 12-14.
- HAUPTMAN, H. (1977). *Acta Cryst.* A33, 553-555.
- HAUPTMAN, H. & KARLE, J. (1953). *Acta Cryst.* 6, 136-141.
- HOWELLS, E. R., PHILLIPS, D. C. & ROGERS, D. (1950). *Acta Cryst.* 3, 210-214.
- KLUG, A. (1958). *Acta Cryst.* 11, 515-543.
- MITRA, G. B. & BELGAUMKAR, J. (1973). *Proc. Indian Nat. Sci. Acad.* 39, 95-100.
- SASS, R. L. & DONOHUE, J. (1958). *Acta Cryst.* 11, 497-504.
- SAYRE, D. (1952). *Acta Cryst.* 5, 60-65.
- SHMUELI, U. (1979). *Acta Cryst.* A35, 282-286.
- SHMUELI, U. (1982). *Acta Cryst.* A38, submitted for publication.
- SHMUELI, U. & WILSON, A. J. C. (1981). *Acta Cryst.* A37, 342-353.
- SIM, G. A. (1958). *Acta Cryst.* 11, 123-124.
- WILSON, A. J. C. (1949). *Acta Cryst.* 2, 318-321.
- WILSON, A. J. C. (1981). *Acta Cryst.* A37, 808-810.

Effects of Heavy Atoms and Symmetry on the Cumulative Distribution Function of Normalised Structure Amplitudes

BY G. D. NIGAM AND SIKHA GHOSH

*Department of Physics, Indian Institute of Technology,
Kharagpur 721302, India*

Abstract

Expansions for the probability density function of the structure factor $|F|$ (or $|E|$), which account for the known contribution of heavy atom in the unit cell have been presented. Using these modified probability density functions, expressions for cumulative distributions $N(z)$ [or $N(|E|)$] have been derived. The expressions are composition and symmetry-dependent and include the effect of atoms in general as well as in fixed special positions. The polynomial series distribution is used to derive an expression for $N(z)$ in the case of a hypercentric crystal. The effect of heavy atoms on two phase structure seminvariants has been estimated in $P\bar{1}$ with the asymptotic form of the distribution. Numerical computation has been carried out to study the effect of heavy atoms, composition and symmetry on $N(z)$. The results have been compared with known crystal structures.

1. Introduction

The probability distribution functions of X-ray intensities were first obtained by Wilson (1949) for space groups $P1$ and $P\bar{1}$. In deriving these distributions, he assumed that the unit cell contained a large number of atoms at random positions and that there was no outstandingly heavy atom in the unit cell. Since then a number of tests based on above distribution functions

have been developed for the verification of space-group assignment and for the resolution of space-group ambiguities. There may be incorrect conclusions if any of the above assumptions are violated. The effects of hypersymmetry (Lipson & Woolfson, 1952; Rogers and Wilson, 1953; Nigam, 1974) and outstandingly heavy atoms (Collin, 1955; Sim, 1958; Foster & Hargreaves, 1963) have been the subject of numerous studies and the literature has been reviewed by Srinivasan & Parthasarathy (1976). In two recent publications (Shmueli, 1979; Shmueli & Wilson, 1981) the cumulative distribution functions $N(|E|)$ [or $N(z)$ —Howells, Phillips & Rogers, 1950] have been generalised to depend on crystallographic symmetry and composition of the asymmetric unit. Suitable probability density functions which were derived for centrosymmetric (Karle & Hauptman, 1953) and non-centrosymmetric (Hauptman & Karle, 1953; Srinivasan & Parthasarathy, 1976) crystals are used. These distribution functions depend on symmetry and atomic heterogeneity of the crystal. Though the effects of heavy atoms in general positions have been considered, heavy atoms in fixed special positions have not been treated explicitly in the analysis.

The present paper is an extension of the works of Shmueli (1979) and Shmueli & Wilson (1981). The probability density functions have been modified in terms of a heavy-atom-dependent parameter $r=f_p/(\Sigma_L)^{1/2}$. The expressions are more general than those obtained by Shmueli (1979), as they include the effects of atoms both in general and in fixed special positions. The polynomial series distribution is used to derive an expression for $N(z)$ in the case of a hypercentric crystal. The effect of heavy atoms has also been studied on two phase structure seminvariants in $P\bar{1}$. Finally, numerical computation is carried out to study the effects of heavy atoms, composition and symmetry. Wherever possible, the results have been tested on a few known crystal structures.

2. Derivation of cumulative distribution functions

Consider a crystal structure of P known heavy atoms in fixed special positions and L light atoms in general positions in the unit cell such that $P+L=N$, where N is the total number of atoms in the unit cell. The structure factor of a reflection \mathbf{h} can be written as

$$F = F_P + F_L, \quad (1)$$

and denote the local average of $|F|^2$, $|F_P|^2$ and $|F_L|^2$ by Σ , Σ_P and Σ_L respectively. That is,

$$\begin{aligned} \langle |F|^2 \rangle &= \sum_{i=1}^N f_{Ni}^2 = \Sigma, \quad \langle |F_P|^2 \rangle = \sum_{i=1}^P f_{Pi}^2 = \Sigma_P, \\ \langle |F_L|^2 \rangle &= \sum_{i=1}^L f_{Li}^2 = \Sigma_L. \end{aligned} \quad (2)$$

The probability that $|F|$ lies between $|F|$ and $|F| + d|F|$ can be derived from the conditional distribution $P(|F|; |F_P|)$ by using the result

$$P(|F|) = \int P(|F|; |F_P|) P(|F_P|) d|F_P|. \quad (3)$$

2.1 Centric case

Since F_L follows the centric distribution, its probability density function can be written as (Karle and Hauptman, 1953)

$$\begin{aligned} P_c(F_L) &= \frac{1}{(2\pi \Sigma_L)^{1/2}} \exp\left(-\frac{F_L^2}{2\Sigma_L}\right) \\ &\times \left[1 + A \left(\frac{1}{3} \frac{F_L^4}{\Sigma_L^2} - 2 \frac{F_L^2}{\Sigma_L} + 1 \right) \right. \\ &\left. + B \left(\frac{1}{15} \frac{F_L^6}{\Sigma_L^3} - \frac{F_L^4}{\Sigma_L^2} + 3 \frac{F_L^2}{\Sigma_L} - 1 \right) + \dots \right], \end{aligned} \quad (4)$$

where

$$A = (\langle |F_L|^4 \rangle_c - 3 \Sigma_L^2) / 8 \Sigma_L^2, \quad (5)$$

$$B = (\langle |F_L|^6 \rangle_c - 15 \Sigma_L \langle |F_L|^4 \rangle_c + 30 \Sigma_L^3) / 48 \Sigma_L^3. \quad (6)$$

From (1) and (4) we obtain the conditional distribution $P_c(|F|; |F_P|)$ to be

$$\begin{aligned} P_c(|F|; |F_P|) = & \frac{1}{\sqrt{2\pi}\Sigma_L} \exp\left(-\frac{(|F| - |F_P|)^2}{2\Sigma_L}\right) \\ & \left[1 + A\left(\frac{1}{3}\frac{(|F| - |F_P|)^4}{\Sigma_L^2} - 2\frac{(|F| - |F_P|)^2}{\Sigma_L} + 1\right) \right. \\ & + B\left(\frac{1}{15}\frac{(|F| - |F_P|)^6}{\Sigma_L^3} - \frac{(|F| - |F_P|)^4}{\Sigma_L^2} \right. \\ & \left. + 3\frac{(|F| - |F_P|)^2}{\Sigma_L} - 1\right) \dots] + \exp\left(-\frac{(|F| + |F_P|)^2}{2\Sigma_L}\right) \\ & \left[1 + A\left(\frac{1}{3}\frac{(|F| + |F_P|)^4}{\Sigma_L^2} - 2\frac{(|F| + |F_P|)^2}{\Sigma_L} + 1\right) \right. \\ & + B\left(\frac{1}{15}\frac{(|F| + |F_P|)^6}{\Sigma_L^3} - \frac{(|F| + |F_P|)^4}{\Sigma_L^2} \right. \\ & \left. + 3\frac{(|F| + |F_P|)^2}{\Sigma_L} - 1\right) \dots]. \quad (7) \end{aligned}$$

Equation (7) can be made more compact when polynomials appearing in it are expressed in terms of Hermite polynomials, $H_n(x)$ using the following identities

$$\frac{1}{3}x^4 - 2x^2 + 1 = \frac{1}{12}H_4\left(\frac{x}{\sqrt{2}}\right), \quad (8)$$

$$\frac{1}{15}x^6 - x^4 + 3x^2 - 1 = \frac{1}{120}H_6\left(\frac{x}{\sqrt{2}}\right), \quad (9)$$

For simplicity, we consider the case of an outstandingly heavy atom at a fixed position and L heterogeneous atoms in general positions in the unit cell. In this case the distribution of F_P is a delta function given by

$$P(F_P) = \delta(F_P - f_P), \quad (10)$$

where f_P is the atomic scattering factor of heavy atom. Using (3), (7), (8) (9) and (10) we obtain $P(|F|)$ d $|F|$ as

$$\begin{aligned} P_c(|F|) d|F| &= \frac{1}{\sqrt{2\pi\Sigma_L}} \left[\exp\left(-\frac{(|F| - f_P)^2}{2\Sigma_L}\right) \right. \\ &\left\{ 1 + \frac{A}{12} H_4\left(\frac{|F| - f_P}{\sqrt{2\Sigma_L}}\right) + \frac{B}{120} H_6\left(\frac{|F| - f_P}{\sqrt{2\Sigma_L}}\right) \dots \right\} + \\ &\exp\left(-\frac{(|F| + f_P)^2}{2\Sigma_L}\right) \left\{ 1 + \frac{A}{12} H_4\left(\frac{|F| + f_P}{\sqrt{2\Sigma_L}}\right) + \frac{B}{120} H_6\left(\frac{|F| + f_P}{\sqrt{2\Sigma_L}}\right) + \dots \right\} \left. \right] d|F|. \end{aligned} \quad (11)$$

Substituting $z = |F|^2/(\Sigma_L + f_P^2)$ and $r = f_P/\Sigma_L^{1/2}$, one gets from (11)

$$\begin{aligned} P_c(z)dz &= \frac{1}{2} \{ (1 + r^2)/2\pi z \}^{1/2} \\ &\times \left[\exp\left(-\frac{[(1+r^2)^{1/2} z^{1/2} - r]^2}{2}\right) \right. \\ &\times \left\{ 1 + \frac{A}{12} H_4\left(\frac{(1+r^2)^{1/2} z^{1/2} - r}{\sqrt{2}}\right) \right. \\ &+ \frac{B}{120} H_6\left(\frac{(1+r^2)^{1/2} z^{1/2} - r}{\sqrt{2}}\right) + \dots \left. \right\} \\ &+ \exp\left(-\frac{[(1+r^2)^{1/2} z^{1/2} + r]^2}{2}\right) \\ &\times \left\{ 1 + \frac{A}{12} H_4\left(\frac{(1+r^2)^{1/2} z^{1/2} + r}{\sqrt{2}}\right) \right. \\ &+ \frac{B}{120} H_6\left(\frac{(1+r^2)^{1/2} z^{1/2} + r}{\sqrt{2}}\right) + \dots \left. \right\} \left. \right]. \end{aligned} \quad (12)$$

The cumulative distribution of z is obtained by computing the integral $\int_0^z P(z') dz'$, which gives $N_c(z) = \int_0^z P(z') dz'$ or in terms of $|E|$ by setting $z = |E|^2$ as

$$\begin{aligned}
 N_c(z) = & \frac{1}{2} \operatorname{erf} \left(\frac{(1+r^2)^{1/2} z^{1/2} + r}{\sqrt{2}} \right) \\
 & - \left(\frac{1}{\pi} \right)^{1/2} \exp \left[- \{ (1+r^2)^{1/2} z^{1/2} + r \}^2 / 2 \right] \\
 & \times \frac{A}{12} H_3 \left(\frac{(1+r^2)^{1/2} z^{1/2} + r}{\sqrt{2}} \right) \\
 & + \frac{B}{120} H_5 \left(\frac{(1+r^2)^{1/2} z^{1/2} + r}{\sqrt{2}} \right) \\
 & + \frac{1}{2} \operatorname{erf} \left(\frac{(1+r^2)^{1/2} z^{1/2} - r}{\sqrt{2}} \right) \\
 & - \left(\frac{1}{\pi} \right)^{1/2} \exp \left[- \{ (1+r^2)^{1/2} z^{1/2} - r \}^2 / 2 \right] \\
 & + \frac{A}{12} H_3 \left(\frac{(1+r^2)^{1/2} z^{1/2} - r}{\sqrt{2}} \right) \\
 & + \frac{B}{120} H_5 \left(\frac{(1+r^2)^{1/2} z^{1/2} - r}{\sqrt{2}} \right) \Big] \dots, \quad (13)
 \end{aligned}$$

$$\begin{aligned}
 N_c(|E|) = & \frac{1}{2} \operatorname{erf} \frac{E - E_p}{\sqrt{2}} - \frac{1}{\sqrt{\pi}} \\
 & \exp - \frac{(E - E_p)^2}{2} - \left[\frac{A}{12} H_3 \left(\frac{E - E_p}{\sqrt{2}} \right) \right. \\
 & \left. + \frac{B}{120} H_5 \left(\frac{E - E_p}{\sqrt{2}} \right) \right] + \frac{1}{2} \operatorname{erf} \frac{E + E_p}{\sqrt{2}} \\
 & - \frac{1}{\sqrt{\pi}} \exp - \frac{(E + E_p)^2}{2} \left[\frac{A}{12} H_3 \left(\frac{E + E_p}{\sqrt{2}} \right) \right. \\
 & \left. + \frac{B}{120} H_5 \left(\frac{E + E_p}{\sqrt{2}} \right) \right], \quad (14)
 \end{aligned}$$

where E and E_p are the normalized structure factors of total and heavy atom components respectively.

2.2. Acentric case

The suitable density function for F_L is given by (Hauptman and Karle, 1953)

$$P_a(F_L) = \frac{1}{\pi \Sigma_L} \exp \left(-\frac{F_L^2}{\Sigma_L} \right) \left[1 + C \left(\frac{1}{2} \frac{F_L^4}{\Sigma_L^2} - 2 \frac{F_L^2}{\Sigma_L} + 1 \right) + D \left(\frac{1}{6} \frac{F_L^6}{\Sigma_L^3} - \frac{3}{2} \frac{F_L^4}{\Sigma_L^2} + 3 \frac{F_L^2}{\Sigma_L} - 1 \right) + \dots \right], \quad (15)$$

where

$$C = (\langle |F_L|^4 \rangle_a - 2 \Sigma_L^2) / 2 \Sigma_L^2, \quad (16)$$

$$D = (\langle |F_L|^6 \rangle_a - 9 \Sigma_L \langle |F_L|^4 \rangle_a + 12 \Sigma_L^3) / 6 \Sigma_L^3. \quad (17)$$

Since F is a vector, we may rewrite (1) as

$$|F_L|^2 = |F|^2 + |F_P|^2 - 2 |F| |F_P| \cos \alpha \quad (18)$$

where α is the angle between F and F_P . From (15) and (18), integrating with respect to α in the range 0 to 2π , $P_a(|F|; |F_P|)$ is obtained. The relation

$$\int_0^{2\pi} \exp(z \cos \alpha) \cos m\alpha \, d\alpha = 2\pi I_m(z), \quad (19)$$

where $I_m(z)$ is the modified Bessel function, is used for integration. From $P_a(|F|; |F_P|)$, after proper simplification for single heavy atom, using (10) and substituting for z , we finally arrive at

$$\begin{aligned} P_a(z) \, dz = & (1+r^2) \exp - \{r^2 + z(1+r^2)\} [I_0(kz^{1/2}) \\ & + C \{(\frac{1}{2}(z(1+r^2) + r^2)^2 + z(1+r^2) \cdot r^2) I_0(kz^{1/2}) \\ & - 2 \{r^2 + (1+r^2)z\} I_0(kz^{1/2}) + I_0(kz^{1/2})\}] \\ & + D \{ \frac{1}{6} [(z(1+r^2) + r^2)^3 + 6(z(1+r^2) \\ & + r^2)(z(1+r^2) \cdot r^2)] I_0(kz^{1/2}) - \frac{3}{2} [(z(1+r^2) + r^2 \\ & + 2z r^2(1+r^2)] I_0(kz^{1/2}) - 3z r^2(1+r^2) I_0(kz^{1/2}) \end{aligned}$$

$$\begin{aligned}
& + 3 (z (1+r^2) + r^2) I_0 (kz^{1/2}) - I_0 (kz^{1/2}) \} \\
& + C \{ -(2z (1+r^2) + r^2) z^{1/2} (1+r^2)^{1/2} \cdot r I_1 (kz^{1/2}) \\
& + zr^2 (1+r^2) I_2 (kz^{1/2}) + 4r^2 z (1+r^2) I_1 (kz^{1/2}) \} \\
& + D \{ -[z^{1/2} r (1+r^2) (z (1+r^2) + r^2)^2 \\
& + (z (1+r^2) + r^2)^3] I_1 (kz^{1/2}) + (z (1+r^2) + r^2) \\
& \times z (1+r^2) r^2 I_2 (kz^{1/2}) - \frac{1}{3} z^{3/2} (1+r^2)^{3/2} r^3 I_3 (kz^{1/2}) \\
& + 6 (z (1+r^2) + r^2) z^{1/2} r (1+r^2) I_1 (kz^{1/2}) \\
& - 6 z^{1/2} r (1+r^2)^{1/2} I_1 (kz^{1/2}), \quad (20)
\end{aligned}$$

where

$$k = 2r (1 + r^2)^{1/2}.$$

The cumulative distribution is obtained in a usual way by numerical technique. We have considered $C_{20}I$, a

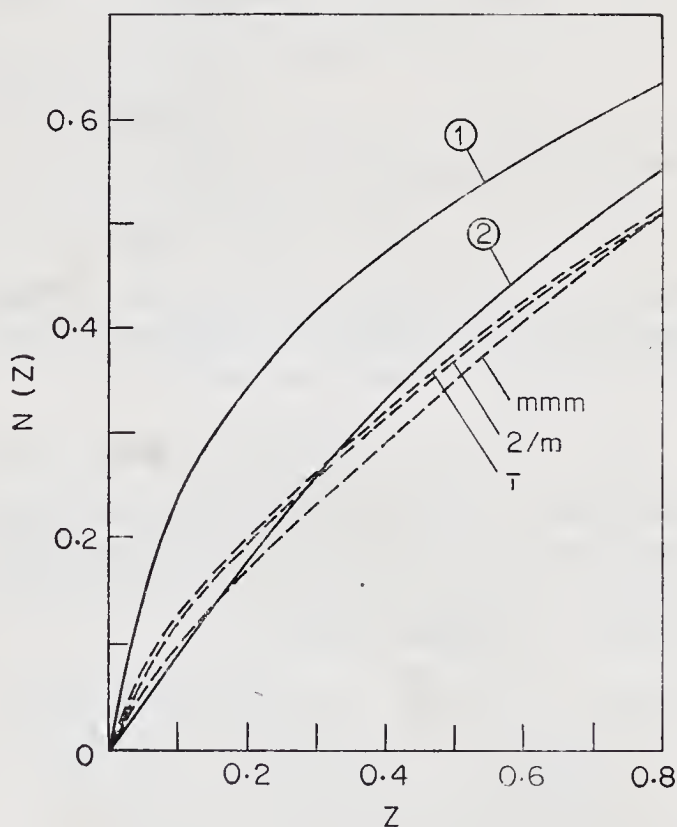


Fig. 1. Theoretical $N(z)$ distribution curves for $\bar{1}$, $2/m$, mmm (dashed lines) compared with Wilson's centric and acentric curves (solid lines) for $C_{20}I$.

hypothetical crystal structure (Shmueli, 1979). The values of A and B for different centrosymmetric point groups are known. The value of r is taken as

$$r = \frac{f_p}{(\Sigma_L)^{1/2}} = \frac{f_I}{(\Sigma_C)^{1/2}} \text{ at } \sin\theta=0.45.$$

The results of calculations are shown in Fig. 1. It is seen that the $N(z)$ functions for different centric point groups as calculated from (13) are more clustered around Wilson-type acentric case. It is consistent with Sim's analysis and may be attributed to the heavy atom effect. Fig. 2 shows the experimental $N(z)$ plot for rubidium di-*o*-nitrobenzoate (Sim, 1958) compared with cumulative distribution function as calculated from (13) and compared with Sim's theoretical values. It is observed that the theoretical curve as calculated from (13) agrees better than that fitted by Sim (1958).

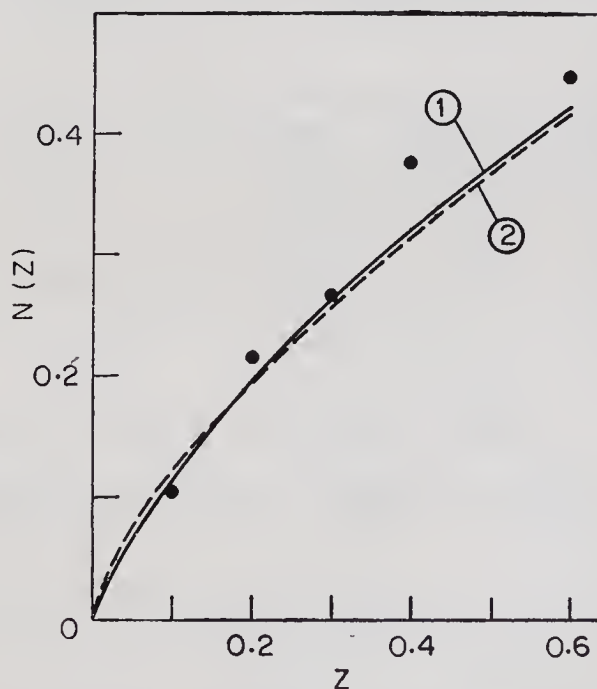


Fig. 2. Comparison of experimental $N(z)$ points (marked by circles) of Rubidium-*o*-dinitrobenzoate with (1) Theoretical distribution from (13) and (2) Sim's curve.

3. Derivation of hypercentric distribution

It is well-known (Wilson, 1949) that pseudosymmetry can give rise to abnormal intensity distribution. The effect has been studied using Wilson statistics. We examine, in what follows, hypercentric intensity distribution using an expansion of probability function as given in (4). It is thus possible to include the effect of symmetry and composition on the corresponding cumulative distribution functions. The hypercentric intensity distribution arises when two molecules, each with an inversion centre occupy general position in a centrosymmetric space group. Following Lipson and Woolfson (1952) and a suitable form of probability density distribution, the probability that F lies between F and $F + dF$ is given by

$$P_h(F) dF = \frac{1}{2} (4\pi\Sigma)^{-1/2} \int_{u=-1}^{u=1} \exp \left(-F^2 \sec^2 \frac{1}{2}\pi u / 4\Sigma \right) \left[1 + AH_4 \left(\frac{F \sec \frac{1}{2}\pi u}{\sqrt{4\Sigma}} \right) + BH_6 \left(\frac{F \sec \frac{1}{2}\pi u}{\sqrt{4\Sigma}} \right) + \dots \right] \sec \frac{1}{2}\pi u dF du, \quad (21)$$

where $u=4s.r'$, s is the reciprocal vector and r' is the position of molecular centre of inversion relative to that of the unit cell. Substituting $t = \tan \frac{1}{2}\pi u$ and $z = F^2/\Sigma$, one gets from (21)

$$P_h(z) dz = \pi^{-1} (4\pi z)^{-1/2} \int_0^{t=\infty} \exp \left\{ -\frac{1}{4}z (1 + t^2) \right\} \times \left[1 + AH_4 \left(\frac{z^{1/2} (1 + t^2)^{1/2}}{2} \right) + BH_6 \left(\frac{z^{1/2} (1 + t^2)^{1/2}}{2} \right) + \dots \right] dt dz. \quad (22)$$

The cumulative distribution function $N_h(z)$ follows from (22) as

$$N_h(z) = \pi^{-1} (4\pi)^{-1/2} \int_0^z \int_0^\infty z^{-1/2} \exp \left\{ -\frac{z}{4} (1 + t^2) \right\} \left[1 + AH_4 \left(\frac{z^{1/2} (1 + t^2)^{1/2}}{2} \right) + BH_6 \left(\frac{z^{1/2} (1 + t^2)^{1/2}}{2} \right) + \dots \right] dt dz.$$

Setting $t = \tan \psi$ and $z = \frac{2\alpha^2}{1+t^2}$, the integration with respect to z can be evaluated and so one gets

$$N_h(z) = \frac{2}{\pi} \int_0^{\pi/2} \operatorname{erf} \left(\frac{\sqrt{z}}{2} \sec \psi \right) d\psi - \frac{4A}{\pi^{3/2}} \int_0^{\pi/2} H_3 \left(\frac{\sqrt{z}}{2} \sec \psi \right) \exp \left(-\frac{z}{4} \sec^2 \psi \right) d\psi - \frac{4B}{\pi^{3/2}} \int_0^{\pi/2} H_5 \left(\frac{\sqrt{z}}{2} \sec \psi \right) \exp \left(-\frac{z}{4} \sec^2 \psi \right) d\psi. \quad (23)$$

The expression for $N_h(|E|)$ follows from (23) by substituting $E^2 = z$. The integrals in (23) were evaluated numerically in order to compare the experimental $N(z)$ distribution (*h o l* projection) data of pyrene ($C_6 H_{10}$) which is an example of hypercentric crystal. Fig.3 shows the results. Fig.4 shows the $N(z)$ plot from (*h k l*) data of 7,7,8-8-tetracyanoquino-dimethane-phenazine complex (Goldberg and Shmueli, 1973). It

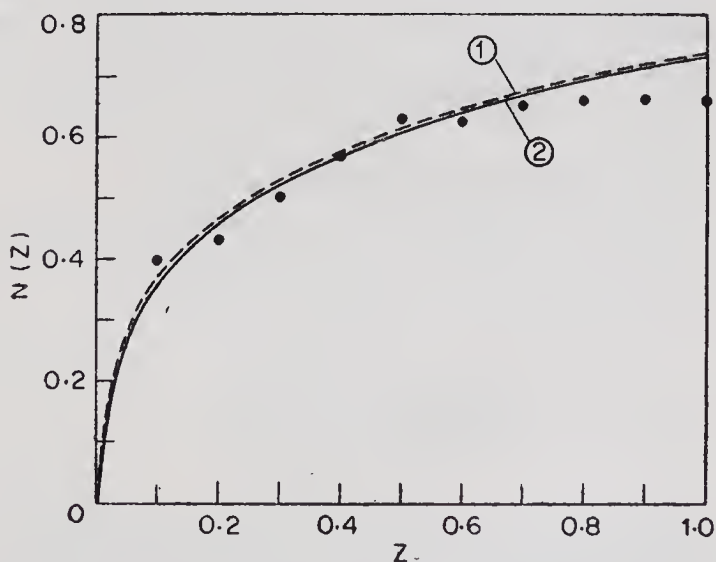


Fig. 3. Comparison of experimental $N(z)$ points (marked by O) of Pyrene (*h o l* projection) with (1) bicentric distribution due to Lipson & Woolfson (1952) and (2) Theoretical distribution from (23).

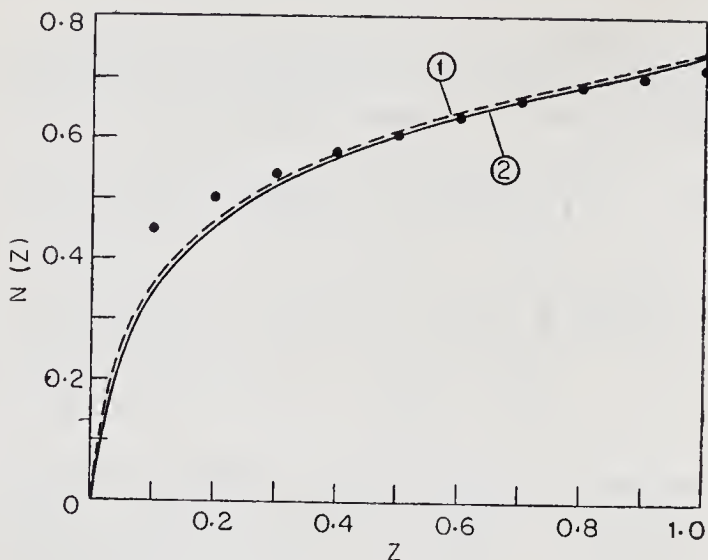


Fig. 4. Comparison of experimental $N(z)$ points (marked by circles) of 7, 7, 8, 8, tetracyanoquinodimethane-phenazine complex with (1) bicentric curve due to Lipson & Woolfson (1952) and (2) Theoretical distribution from (23).

is seen that agreement in both the cases is fairly good. The theoretical curve shows a slight tendency to be displaced towards Wilson's centric distribution. This may be attributed to the inclusion of higher order terms in the polynomial density distribution function.

4. Heavy-atom contribution and two phase structure seminvariants in space group $P\bar{1}$

In space group $P\bar{1}$, the linear combination

$$\psi = \phi_{\mathbf{h}} + \phi_{\mathbf{k}}, \quad (24)$$

is a structure seminvariant if and only if

$$\mathbf{h} + \mathbf{k} = 0 \bmod (\boldsymbol{\omega}), \quad (25)$$

where $\boldsymbol{\omega} = (2, 2, 2)$

The normalized structure factor of reflections $\mathbf{h}(h_1 k_1 l_1)$ and $\mathbf{k}(h_2 k_2 l_2)$ can be written as

$$\Sigma^{1/2} E(\mathbf{h}) = \Sigma_P^{1/2} E_P(\mathbf{h}) + \Sigma_L^{1/2} E_L(\mathbf{h}), \quad (26)$$

$$\Sigma^{1/2} E(\mathbf{k}) = \Sigma_P^{1/2} E_P(\mathbf{k}) + \Sigma_L^{1/2} E_L(\mathbf{k}) \quad (27)$$

where E_P and E_L are the normalized structure factors of the heavy-atom component and the light-atom component respectively. The number of light atoms is assumed to be large so that Wilson's centric distribution is valid. The joint probability distribution $P(E(\mathbf{h}), E(\mathbf{k}), E_P(\mathbf{h}), E_P(\mathbf{k}))$ can be written as (Klug, 1958)

$$P(E(\mathbf{h}), E(\mathbf{k}), E_P(\mathbf{h}), E_P(\mathbf{k})) = \frac{1}{2\pi} \exp -\frac{1}{2} [(E(\mathbf{h}) - E_P(\mathbf{h}))^2 + (E(\mathbf{k}) - E_P(\mathbf{k}))^2] \quad (28)$$

The conditional joint probability distribution $P(E(\mathbf{h}), E(\mathbf{k}); E_P(\mathbf{h}), E_P(\mathbf{k}))$ can be readily obtained from the relation

$$P(E(\mathbf{h}), E(\mathbf{k}); E_P(\mathbf{h}), E_P(\mathbf{k})) = \frac{P(E(\mathbf{h}), E(\mathbf{k}), E_P(\mathbf{h}), E_P(\mathbf{k}))}{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} P(E(\mathbf{h}), E(\mathbf{k}), E_P(\mathbf{h}), E_P(\mathbf{k})) dE(\mathbf{h}) dE(\mathbf{k})} \quad (29)$$

From (28) and (29) one gets

$$P(E(\mathbf{h}), E(\mathbf{k}); E_P(\mathbf{h}), E_P(\mathbf{k})) = \frac{1}{2\pi} \exp -\frac{1}{2} [(E(\mathbf{h}) - E_P(\mathbf{h}))^2 + (E(\mathbf{k}) - E_P(\mathbf{k}))^2]. \quad (30)$$

Equation (30) may be used to calculate $\langle E(\mathbf{h})E(\mathbf{k}); E_P(\mathbf{h})E_P(\mathbf{k}) \rangle$ as $\langle E(\mathbf{h})E(\mathbf{k}); E_P(\mathbf{h})E_P(\mathbf{k}) \rangle$

$$= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} E(\mathbf{h})E(\mathbf{k}) P(E(\mathbf{h})E(\mathbf{k}); E_P(\mathbf{h})E_P(\mathbf{k})) dE(\mathbf{h})dE(\mathbf{k}) = E_P(\mathbf{h})E_P(\mathbf{k}). \quad (31)$$

Similarly

$$\begin{aligned} \langle E^2(\mathbf{h})E^2(\mathbf{k}); E_P(\mathbf{h})E_P(\mathbf{k}) \rangle &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} E^2(\mathbf{h})E^2(\mathbf{k}) \\ P(E(\mathbf{h})E(\mathbf{k}); E_P(\mathbf{h})E_P(\mathbf{k})) dE(\mathbf{h})dE(\mathbf{k}) &= 1 + E_P^2(\mathbf{h}) + E_P^2(\mathbf{k}) \\ + E_P^2(\mathbf{h})E_P^2(\mathbf{k}). \end{aligned} \quad (32)$$

We may now expand the conditional probability

distribution of the random variable $R=E(\mathbf{h})E(\mathbf{k})$ in Gram-Charlier series (Cramér, 1951): we obtain

$$P(R;E_P(\mathbf{h}), E_P(\mathbf{k})) = \frac{1}{2\pi\sigma^2} \exp \left[-\frac{(R-\langle R \rangle)^2}{2\sigma^2} \right] + \dots \quad (33)$$

where $\langle R \rangle$ is given by (31), and

$$\sigma^2 = 1 + E_P^2(\mathbf{h}) + E_P^2(\mathbf{k}). \quad (34)$$

If one denotes the probability that P_+ has the same sign as $E_P(\mathbf{h})E_P(\mathbf{k})$, then it can readily be shown that (Klug, 1958; Giacovazzo, 1975)

$$P_+ = \frac{1}{2} + \frac{1}{2} \tanh \frac{|E(\mathbf{h})E(\mathbf{k})| |E_P(\mathbf{h})E_P(\mathbf{k})|}{1 + E_P^2(\mathbf{h}) + E_P^2(\mathbf{k})} \quad (35)$$

Equation (35) can be used to estimate the sign of the product $E(\mathbf{h})E(\mathbf{k})$. This formula fulfils the requirement that $P_+ = \frac{1}{2}$ whenever $E_P(\mathbf{h})=0$ or $E_P(\mathbf{k})=0$. Equation (35) was tested on $hk0$ data of metanilic acid (Hall & Maslen, 1965). In this structure $\Sigma f_{\text{light}}^2 > \Sigma f_{\text{heavy}}^2$, thus indicating that the contribution from heavy atoms is not large enough to dominate the phases of the structure. The results of calculations are presented in Table 1. It is observed that whenever P_+ is large ($P_+ \geq 0.7$) the product $E(\mathbf{h})E(\mathbf{k})$ has the same sign as that of $E_P(\mathbf{h})E_P(\mathbf{k})$. It is further noted that the probabilities of the sign of the $E(\mathbf{h})E(\mathbf{k})$ are always dominated by heavy atom contribution to the normalized structure factors and the probabilities are high for large products. The probabilities of sign indications are always poor when the structure factors involved are small in magnitude.

5. Conclusions

The present study confirms that the modified probability density functions are well suited for the evaluation of $N(|E|)$ or $N(z)$ functions, as the effects of heavy

Table 1. *Probability calculations for signs of reflections for metanilic acid*

$E(\mathbf{h})$	$E(\mathbf{k})$	Sign of $E(\mathbf{h})$ $E(\mathbf{k})$ observed	P_+ from (35)	Sign of $E_P(\mathbf{h})$ $E_P(\mathbf{k})$
$E_{1,1}$	$E_{1,3}$	+	0.57	—
$E_{1,3}$	$E_{1,5}$	+	0.70	+
$E_{1,5}$	$E_{1,7}$	+	0.66	—
$E_{1,7}$	$E_{1,9}$	—	0.62	+
$E_{1,9}$	$E_{1,11}$	+	0.57	—
$E_{1,11}$	$E_{1,13}$	—	0.54	+
$E_{2,8}$	$E_{2,10}$	+	0.55	—
$E_{2,12}$	$E_{2,14}$	+	0.70	+
$E_{5,5}$	$E_{5,7}$	+	0.99	+
$E_{5,9}$	$E_{5,11}$	—	0.99	—
$E_{3,1}$	$E_{3,3}$	—	0.85	—
$E_{3,11}$	$E_{3,13}$	+	0.75	+

atoms, symmetry and composition of the asymmetric unit are included in the final expressions. In case of light-atom hypersymmetric structures, the cumulative distribution functions given by Lipson and Woolfson (1952) and Rogers and Wilson (1953) are accurate enough to detect the hypersymmetry. The effect of higher-order terms is not very significant, but the situation may change altogether if hypersymmetry and an outstandingly heavy atom are present simultaneously.

The formula (35) may be used together with known structural information to compute the phases of structure seminvariants which are essential in any phase-determination process by direct methods. It would be interesting to have a formula for the sign of structure seminvariants in which both structural information and neighbourhood concept are included simultaneously. This is the subject of further investigation.

References

- COLLIN, R. L. (1955). *Acta Cryst.* **8**, 499–502.
- CRAMÉR, H. (1951). *Mathematical Methods of Statistics*, Princeton University Press.
- FOSTER, F. & HARGREAVES, A. (1963). *Acta Cryst.* **16**, 1124–1133.
- GIACOVAZZO, G. (1975). *Acta Cryst.* **A31**, 252–259.
- GOLDBERG, I. & SHMUELI, U. (1973) *Acta Cryst.* **B29**, 440–448.
- HALL, S. R. & MASLEN, E. N. (1965). *Acta Cryst.* **18**, 301–306.
- HAUPTMAN, H. & KARLE, J. (1953). *Acta Cryst.* **6**, 136–141.
- HOWELLS, E. R., PHILLIPS, D. C. & ROGERS, D. (1950). *Acta Cryst.* **3**, 210–214.
- KARLE, J. & HAUPTMAN, H. (1953). *Acta Cryst.* **6**, 131–135.
- KLUG, A. (1958). *Acta Cryst.* **11**, 515–543.
- LIPSON, H. & WOOLFSON, M. M. (1952). *Acta Cryst.* **5**, 680–682.
- NIGAM, G. D. (1974). *Z. Kristallogr.* **140**, 336–343.
- ROGERS, D. & WILSON, A. J. C. (1953), *Acta Cryst.* **6**, 439–449.
- SHMUELI, U. (1979). *Acta Cryst.* **A35**, 282–286.
- SHMUELI, U. & WILSON, A. J. C. (1981). *Acta Cryst.* **A37**, 342–353.
- SIM, G. A. (1958). *Acta Cryst.* **11**, 123–124.
- SRINIVASAN, R. & PARTHASARATHY, S. (1976). *Some Statistical Applications in X-Ray Crystallography*, Oxford: Pergamon Press.
- WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.

Measurability of Bijvoet Differences*

BY S. PARTHASARATHY

*Department of Crystallography and Biophysics,
University of Madras, Guindy Campus,
Madras 600 025, India*

Abstract

The determination of the absolute configuration of molecules and structure determination of non-centrosymmetric (NC) crystals with heavy atoms are two of the important applications of anomalous scattering with phase shift. A statistical method of selecting the few optimum reflections for Bijvoet difference (BD) measurement (at the stage where the positions of the heavy atoms are known) for the purpose of determining the absolute configuration is pointed out. The success of the anomalous scattering method for determining the structures of NC crystals depends on the possibility of measuring accurately the BDs of a large percentage of reflections. The measurability of BDs can be studied from a knowledge of the probability distribution of normalized BD variables or the Bijvoet ratio. The measurability is expected to be influenced by structural features (*e.g.* presence of centrosymmetric parts in the molecule, space-group symmetry and the degree of centrosymmetry of the crystal) as well as the non-observability of extremely weak reflections. After dealing with the optimum conditions for observing large BDs in a perfectly NC crystal, the influence of various structural features and of data truncation on the measurability are considered.

1. Introduction

If the wavelength of incident *X*-rays is close to and

*Contribution No. 581.

less than the absorption edge of an atom the scattering from the atom becomes anomalous such that the scattering factor has a component which is 90° ahead in phase with respect to normal scattering. When there is anomalous scattering with phase shift (see Ramaseshan, 1964 for the terminology) the intensities of the inverse reflections $\mathbf{H} (=h\ k\ l)$ and $\bar{\mathbf{H}} (= \bar{h}\ \bar{k}\ \bar{l})$ of a non-centrosymmetric (NC, hereafter) crystal are no more equal resulting in the breakdown of Friedel's law (James, 1958). The difference in the intensities of the inverse reflections is called the Bijvoet difference (BD, hereafter) owing to the pioneering work of Bijvoet on the crystallographic applications of this difference (Bijvoet, 1952, 1954, 1955). Two of the important uses of BD measurement are: (i) determining phases of reflections in NC crystals and using them to elucidate the structure (Ramachandran & Raman, 1956; Peerdeman & Bijvoet, 1956) and (ii) establishing the absolute configuration of molecules or atomic arrangement in NC crystals (Bijvoet, Peerdeman & van Bommel, 1951). While structure determination of an NC crystal by the anomalous scattering method* *via* either the quasi-anomalous synthesis (Ramachandran & Raman, 1956; Ramachandran & Parthasarathy, 1965) or the weighted anomalous synthesis (Parthasarathy, Ramachandran & Srinivasan, 1964; Sim, 1964) requires the accurate measurement of the BDs of a fairly large percentage of reflections, the determination of the absolute configuration by the Bijvoet method requires the measurement of BDs of a few (a dozen, say) reflections only. With respect to these applications we shall use the concepts of 'measurability of BDs of a

*We use the term 'structure determination by anomalous scattering method' to mean the determination of the positions of a sufficient percentage of atoms in the structure from either the quasi-anomalous or weighted anomalous synthesis. The rest of the atoms in the unit cell can then be determined by the standard Fourier-methods (see Ramachandran & Srinivasan, 1970).

crystal' and 'measurability of BD of a reflection'. The measurability of BDs of a crystal is concerned with the determination of the suitability of a crystal for BD data collection for the purpose of structure determination while the measurability of BD of a reflection is concerned with the determination of the suitability of any given reflection for BD measurement. In this article we shall define the two measurabilities quantitatively, using probability criteria. We shall then discuss the optimum conditions for the measurability of BDs of a crystal and study how this measurability is influenced by a number of structural and non-structural features. We shall also discuss how the concept of measurability of BD of a reflection can be used for selecting the few optimum reflections for BD measurement for the purpose of establishing the absolute configuration.

2. Notation and preliminary results

Consider an NC crystal containing N atoms in the unit cell of which P atoms are anomalous scatterers and the remaining $Q(=N-P)$ atoms are normal scatterers. In the case of X -rays the anomalous scatterers are generally heavy atoms while the normal scatterers are light atoms such as C, N and O. In our discussion we shall take all the anomalous scatterers in the asymmetric unit to be of the same type and the normal scatterers to be of similar scattering power. Heavy atom derivatives of most organic and biomolecules come under this situation.

The scattering factor of an atom under anomalous scattering with phase shift is a complex quantity and can be written as

$$f = f_0 + f' + i f'', \quad (1)$$

where f_0 is the high frequency limit of the scattering factor and f' and f'' are the real and imaginary dispersion corrections (for this notation see Ramaseshan

& Abrahams, 1975). In the theoretical probability distribution functions of BD and Bijvoet ratio (BR, hereafter) the ratio of the imaginary to the total real part of the atomic scattering factor of the anomalous scatterer enters as one of the parameters of the distributions. We shall hence denote it by k . That is

$$k = f''/(f_0 + f'). \quad (2)$$

The structure factor of a reflection \mathbf{H} can be written in terms of the contributions from the P - and Q -atoms as

$$\begin{aligned} F_N(\mathbf{H}) &= F'_P(\mathbf{H}) + F''_P(\mathbf{H}) + F_Q(\mathbf{H}) \\ &= F'_N(\mathbf{H}) + F''_P(\mathbf{H}), \end{aligned} \quad (3)$$

where

$$F'_N(\mathbf{H}) = F'_P(\mathbf{H}) + F_Q(\mathbf{H}). \quad (4)$$

$F'_P(\mathbf{H})$ and $F''_P(\mathbf{H})$ are the contributions to the structure factor of reflection \mathbf{H} from the real and imaginary parts respectively of the atomic scattering factor of the P -atoms and $F_Q(\mathbf{H})$ is the contribution to the structure factor from the Q -atoms. The structure factor relations for the inverse reflections \mathbf{H} and $\bar{\mathbf{H}}$ in the presence of anomalous scattering can be conveniently represented in the Argand diagram (Fig. 1). Here the diagram for reflection $\bar{\mathbf{H}}$ is shown reflected about the real axis in order to clearly show the occurrence of the BD. In crystals with a single species of anomalous scatterer we can show that

$$F''_P(\mathbf{H}) = i k F'_P(\mathbf{H}), \quad (5)$$

which implies that

$$|F''_P(\mathbf{H})| = k |F'_P(\mathbf{H})|, \quad \alpha''_P(\mathbf{H}) = \alpha'_P(\mathbf{H}) + \frac{\pi}{2}. \quad (6)$$

We shall define

$$\theta = \alpha'_N - \alpha'_P, \quad \psi = \alpha_Q - \alpha'_P. \quad (7)$$

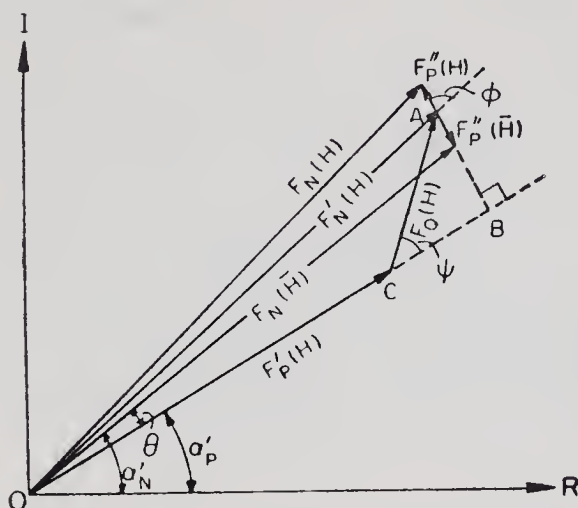


Fig. 1. Argand diagram representation of the structure factor relations for the inverse reflections \mathbf{H} and $\bar{\mathbf{H}}$.

Using geometrical considerations theoretical expressions for the BD and the mean intensity of the inverse reflections \mathbf{H} and $\bar{\mathbf{H}}$ can be derived from Fig. 1 as

$$\Delta I = I(\mathbf{H}) - I(\bar{\mathbf{H}}) = 4 |F'_N| |F''_P| \sin \theta, \quad (8)$$

$$\bar{I} = \frac{1}{2} [I(\mathbf{H}) + I(\bar{\mathbf{H}})] = |F'_N|^2 + |F''_P|^2 \quad (9)$$

where (8) is the Ramachandran-Raman-Peerdeman-Bijvoet formula for BD and is the starting point for the theoretical derivation of the probability distributions of the BD variables for different situations.

2.1. Normalized Bijvoet difference and Bijvoet ratio variables

In the theoretical studies on the measurability of BDs a number of normalized BD and BR variables have been defined and used. The normalized BDs x and Δ are defined to be (Parthasarathy & Srinivasan, 1964; Parthasarathy, 1967).

$$x = \frac{|\Delta I|}{4[\langle |F_Q|^2 \rangle \langle |F''_P|^2 \rangle]^{1/2}}, \quad (10)$$

$$\Delta = |\Delta I| / \langle |F'_N|^2 \rangle. \quad (11)$$

The Bijvoet ratio X is defined to be (Zachariasen, 1965)

$$X = \frac{|\Delta I|}{\frac{1}{2}[I(H) + I(\bar{H})]} = \frac{|\Delta I|}{\bar{I}} = \frac{4 |F'_N| |F''_P| |\sin \theta|}{|F'_N|^2 + |F''_P|^2} \quad (12)$$

For X-rays, in most cases $|F''_P|^2 \ll |F'_N|^2$ so that $|F''_P|^2$ in the denominator of (12) may be neglected (particularly when k is not large) in comparison with $|F'_N|^2$ (Parthasarathy, 1967). We denote the resulting quantity by δ and call it the modified BR. That is

$$\delta = |\Delta I| / |F'_N|^2 = 4 |F''_P| |\sin \theta| / |F'_N|. \quad (13)$$

For the theoretical treatment it is convenient to express these in terms of the normalized structure factor magnitudes y'_N , y'_P and y_Q which are defined as

$$y_Q = |F_Q| / \langle |F_Q|^2 \rangle^{1/2},$$

$$y'_\alpha = |F'_\alpha| / \langle |F'_\alpha|^2 \rangle^{1/2}, \quad \alpha = N \text{ or } P. \quad (14)$$

From (8)–(14) we can readily show that

$$x = y'_P y_Q |\sin \psi|, \quad (15)$$

$$\Delta = 4k\sigma_1\sigma_2x, \quad (16)$$

$$\delta = 4k\sigma_1y'_P |\sin \theta| / y'_N, \quad (17)$$

$$X = \frac{4k\sigma_1y'_Ny'_P |\sin \theta|}{y'^2_N + k^2\sigma_1^2y'^2_P}, \quad (18)$$

where σ_1^2 and σ_2^2 are defined to be

$$\sigma_1^2 = \langle |F'_P|^2 \rangle / \langle |F'_N|^2 \rangle, \quad \sigma_2^2 = \langle |F_Q|^2 \rangle / \langle |F'_N|^2 \rangle. \quad (19)$$

It is useful to note here that though the BD and BR of a reflection can take both positive and negative values, we have defined the variables x , Δ , X and δ to be positive. This is because, for studying the measurability the relevant quantities are only the magnitudes of these variables.

2.2. Nomenclature

In our study we use the terms one-atom, two-atom, many-atom ($P=MN$) and many-atom ($P=MC$) cases in the following sense: Consider a crystal of space group $P1$ containing P heavy and Q light atoms in the unit cell. The situation where $P=1$ is called the one-atom case and that where $P=2$ the two-atom case. When P is many, the P -group can take up either an NC or a centrosymmetric (**C**, hereafter) configuration and these two situations are called the many-atom ($P=MN$) case and many-atom ($P=MC$) case respectively.

3. Measurability of Bijvoet differences of a crystal

3.1. Factors influencing the measurability*

The success of the X-ray anomalous scattering method for structure determination *via* Fourier methods strongly depends on the measurability of BDs, that is, on the possibility of measuring fairly accurately the BDs of a large percentage of reflections. The factors which are expected to influence the measurability can be grouped into two classes, namely, (1) structural and (2) non-structural. Examples for the former are: (i) space-group symmetry, (ii) presence of a centrosymmetric part in a molecule [*i.e.* the degree of centrosymmetry (DCS, hereafter) of the molecule], (iii) DCS of the crystal as a whole, (iv) the presence of pseudosymmetry in the atomic arrangement in the crystal structure, etc. All these structural features may not coexist in a particular crystal structure.

One of the non-structural factors which may be expected to affect the measurability arises from the limitation of physical measurements. It is known that in a crystal not *all* the theoretically possible reflections

*In §3 we shall use the term 'measurability' to stand for 'the measurability of BDs of a crystal'.

in a given $(\sin\theta/\lambda)$ -range* can actually be observed. There always exists a finite percentage of reflections which are too weak to be observed. Owing to the non-observability of extremely weak reflections the observed data will suffer a truncation at the lower end. Such a truncation could affect the measurability. A study of this aspect is also important since large BRs have been generally believed to occur among extremely weak reflections (Ramachandran & Srinivasan, 1970). The other non-structural factor affecting the measurability is the wavelength of the radiation used for data collection. It is obvious that the closeness of the incident wavelength to the absorption edge of an atom could result in enhanced anomalous scattering with a consequent increase in the measurability.

3.2. Bijvoet difference variables used for studying the measurability

The measurability of BDs of a crystal can be studied by obtaining the probability distribution of either the BR X or the normalized BD variables x or Δ . Of these the BR is the best for such studies. Owing to the somewhat complicated functional dependence of X on θ , y'_N and y'_P the distribution function of X cannot be obtained in a closed form. In some of our studies we have therefore made use of the modified BR δ and this to some extent helped to circumvent the theoretical difficulties. However, it is important to note here that the results on measurability obtained from a study of the probability distribution of δ could be somewhat an overestimation of the effect, particularly when the anomalous scattering is quite pronounced (*i.e.* when k is large). In our early papers we used only the variables x and Δ in order to obtain the theoretical distribution functions

*We shall denote $\sin\theta/\lambda$ by S and the maximum value of S for the data by S_{\max} . Here θ stands for the Bragg angle and is different from that in (7).

in closed form and this in turn made the evaluation of these probabilities by manual computation easy.

3.3. Definition of the measurability

A probability measure that is suitable for expressing the measurability of BDs of a crystal is the complementary cumulative function (CCF hereafter) of the BR* and this is denoted by $N_X^c(X_0)$ where X_0 is a particular value of the random variable X . That is,

$$N_X^c(X_0) = \Pr(X \geq X_0), \quad (20)$$

which denotes the probability that the BR X takes a value greater than any specific value X_0 . Physically $N_X^c(X_0)$ represents the fractional number of reflections for which the magnitude of the BR is greater than any given value X_0 . Since a BR which is of the order of 0.1 (*i.e.* 10%) can be easily measured we shall treat $N_X^c(0.1)$ and 'measurability' as synonymous in our discussion.

3.4. Parameters characterizing the measurability

From (18) it is clear that the CCF of X will depend on the parameters k and σ_1^2 . Since X is a function of y'_N , y'_P and θ (see (18)) the probability distribution of X can be obtained from the joint probability distribution function of y'_N , y'_P and θ . As y'_P depends on the number of P -atoms in the asymmetric unit and their configuration the CCF of X may be expected to depend on these characteristics of the P -atoms. Further in crystals with special structural features the specific parameter that characterizes the structural feature will also be a parameter of the CCF of X . Thus though the measurability depends on k , σ_1^2 , the parameter defining special structural feature, etc.,

*The CCF of Δ , namely, $N_\Delta^c(\Delta_0)$ can also be used for this purpose.

the relative importance of these factors can be determined only from a study of the behaviour of the CCF of X with respect to variations in these parameters. We shall take up this aspect in § 3.5. (d).

Of these parameters k is determined by the type of heavy atom and the wavelength of X-rays and hence it is suitable for characterizing the influence of the wavelength on measurability. σ_1^2 is determined by the number of P -atoms and Q -atoms and their scattering powers. σ_1^2 is a convenient measure of the relative domination of the anomalous scatterers in contributing to the local mean intensity. The quantities k and σ_1^2 are not present as parameters of the probability distribution of the variable x and hence it is not as suitable as X for studying the measurability.

3.5. *Measurability study in space group $P1$ when the Q -group is ideally non-centrosymmetric*

Before going into details of how the various factors individually influence the measurability we shall study the influence of the number of anomalous scatterers in the unit cell on measurability, the relative importance of k and σ_1^2 and the optimum conditions for the measurability for crystals of space group $P1$ with all atoms occurring at random positions. That is, the Q -group is assumed to satisfy the requirements of the acentric Wilson distribution (Wilson, 1949).

(a) *The influence of the number of anomalous scatterers.* The CCFs of the normalized BD x are shown in Fig.2 for the cases $P=1,2$, MN and MC (Parthasarathy & Srinivasan, 1964). It is seen that the one-atom case is the most favourable while the many-atom ($P = MC$) case is the least favourable for BD measurement, the other conditions such as k and σ_1^2 being the same. The two-atom and many-atom ($P=MN$) cases are more or less equally effective and fall somewhat between the one-atom and many-atom ($P = MC$) cases.

Values of $N_\delta^c(0.1)$ as a function of k and σ_1^2 are

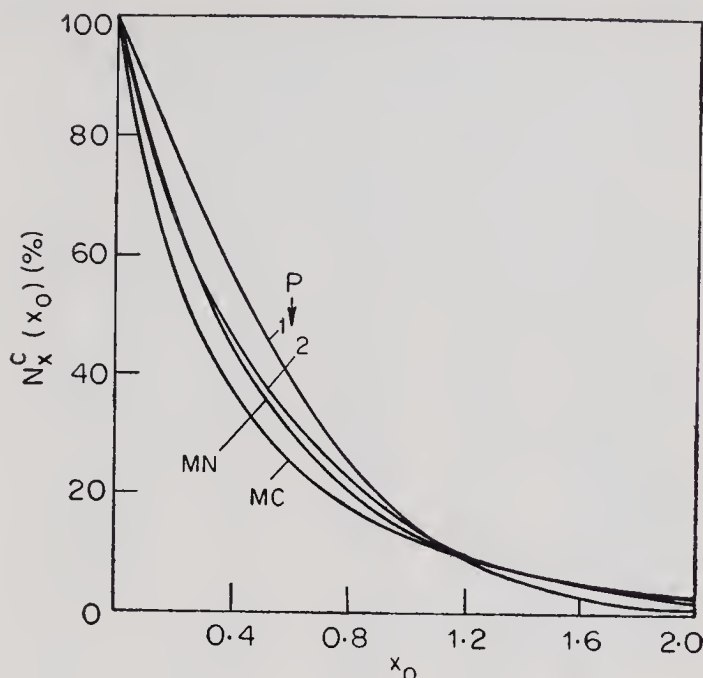


Fig. 2. Complementary cumulative function (in %) of x for the one-atom, two-atom and many-atom ($P = MN$ and $P = MC$) cases.

given in Table 1 for the cases $P=1$, MN and MC (Parthasarathy & Parthasarathi, 1973). It is seen from Table 1 that for given k and σ_1^2 in the region of general interest (*i.e.* $\sigma_1^2 < 0.7$)

$$[N_\delta^c(0.1)]_1 > [N_\delta^c(0.1)]_{MN} > [N_\delta^c(0.1)]_{MC}. \quad (21)$$

Thus, for the given values of k and σ_1^2 , the measurability is the largest in the one-atom case and the least in the many-atom ($P=MC$) case. The many-atom ($P=MN$) case falls somewhat between these two. This is in agreement with the conclusion obtained from a study of CCF of x . It may also be noted that even for the least favourable case (*i.e.* $P=MC$ case), if k and σ_1^2 are not very small, enough percentage of reflections have measurable BDs. For example, corresponding to $k=0.1$ and $\sigma_1^2=0.3$ about 43% of reflections have $\delta > 0.1$. Thus structure determination

Table 1. $N_{\delta}^c(0.1)$ (in %) as a function of k and σ_1^2 for the cases $P=1$, MN and MC

k	P	σ_1^2				
		0.1 $\frac{1}{2}$	0.3	0.5	0.7	0.9
0.06	1	23.0	42.8	43.3	34.0	10.5
	MN	18.8	32.7	36.0	32.7	18.8
	MC	16.2	26.6	29.5	28.2	20.1
0.1	1	46.1	62.6	61.4	53.5	29.0
	MN	36.0	52.1	55.3	52.1	36.0
	MC	30.1	43.4	47.2	46.4	36.4
0.2	1	73.9	80.4	79.4	74.7	58.1
	MN	61.5	73.7	75.7	73.7	61.5
	MC	51.7	64.3	68.1	68.3	60.6
0.3	1	82.7	86.8	86.1	82.8	71.1
	MN	73.2	82.1	83.6	82.1	73.2
	MC	62.9	73.6	76.9	77.5	72.0

could be carried out even for the case $P=MC$ by a proper choice of k and σ_1^2 .

(b) *Relative importance of k and σ_1^2 .* In order to determine the relative importance of k and σ_1^2 for the measurability we shall study (for a fixed P) the variation of $N_{\delta}^c(0.1)$ as a function of σ_1^2 (keeping k constant) and as a function of k (keeping σ_1^2 constant). Such a study enables us to decide on the type of heavy atom to be used for preparing the heavy-atom derivative of a given compound and on the proper X-radiation to be employed for data collection in order to optimize measurability.

The curves of $N_{\delta}^c(0.1)$ vs σ_1^2 for different fixed values of k are shown in Fig. 3 for the case $P=MN$. It is seen that as σ_1^2 increases the percentage of reflections for which $\delta > 0.1$ increases and attains a maximum at $\sigma_1^2 = 0.5$, thereafter decreasing to zero at $\sigma_1^2 = 1.0$. Thus, for a given k , the measurability will be the largest when σ_1^2 is close to 0.5 and least when it is either 0 or 1. Thus a P -group whose relative domi-

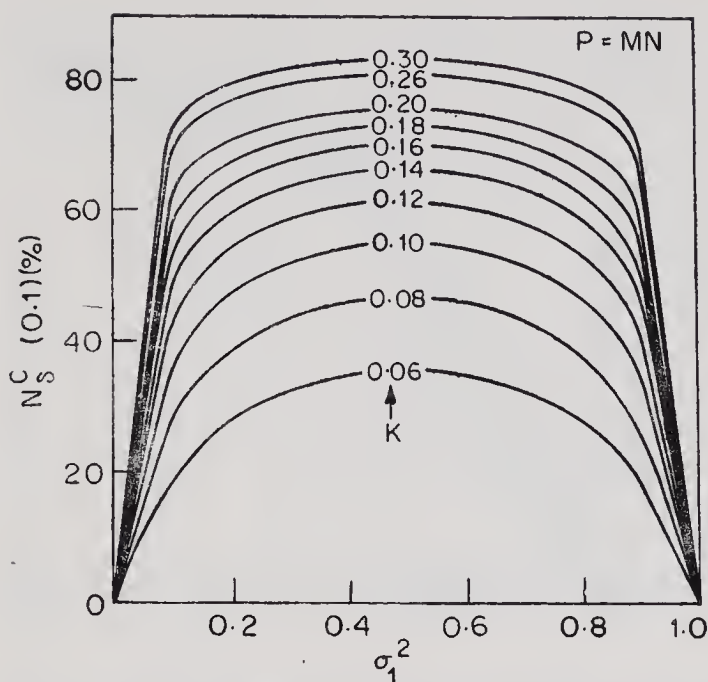


Fig. 3. Complementary cumulative function (in %) of δ as a function of σ_1^2 for different fixed values of k for the many-atom ($P = MN$) case. The number on each curve denotes the value of k .

nation is either too much ($\sigma_1^2 > 0.9$, say) or too little ($\sigma_1^2 < 0.1$, say) will not be suitable for optimum BD measurement.

The variation of $N_s^c(0.1)$ as a function of k for different fixed values of σ_1^2 is shown in Fig. 4 for the case $P = MN$. It is seen that the percentage of reflections for which $\delta > 0.1$ increases systematically as k increases. Thus, for a given σ_1^2 measurability increases as k increases. This shows that for realizing the full power of the anomalous scattering method in structure analysis the choice of proper wavelength for data collection is of great importance.

(c) *Optimum conditions.* From the study of the nature of $N_s^c(0.1)$ as functions of k and σ_1^2 (see §§ 3.5 (a), (b)) we obtain the following as optimum conditions for the measurability in crystals of space group $P1$:

C. S.—10

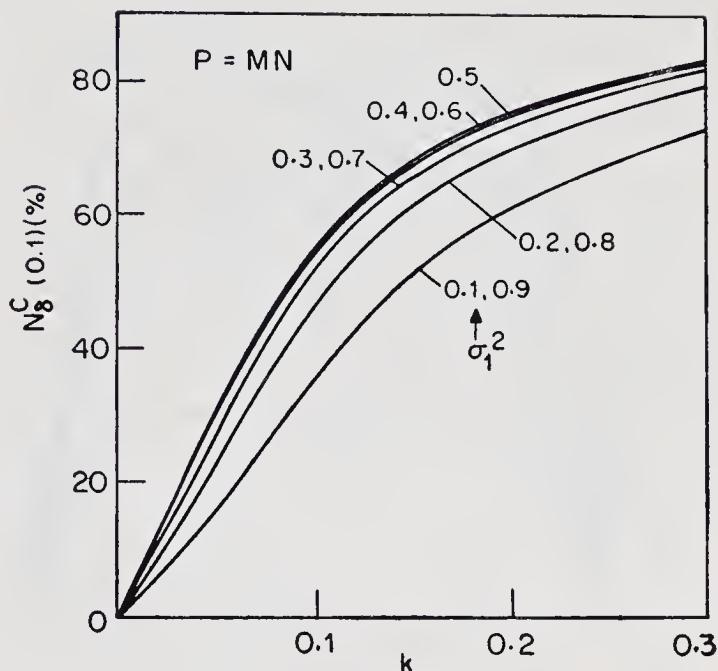


Fig. 4. Complementary cumulative function (in %) of δ as a function of k for different fixed values of σ_1^2 for the many-atom ($P = MN$) case. The numbers near the curves denote the values of σ_1^2 .

(i) k should be as large as possible*, (ii) σ_1^2 should be close to 0.5 and (iii) the number of anomalous scatterers in the unit cell should be one. It may be noted here that conditions (i) and (ii) are also found to hold good in space groups of higher symmetry (see § 3.5. (j)).

(d) *Effect of data-truncation.* Suppose y_t is the threshold value of the normalized structure factor magnitude for the data. That is, the reflections for which $y'_N < y_t$ are taken to be too weak to be observed.

*The measurability of BDs in the case of Factor V 1a (Dale *et al.* 1963), Methyl melaleucate iodoacetate (Hall & Maslen, 1965) and Davallol iodoacetate (Yow-Lam Oh & Maslen 1966) were so large as to enable these structures determined from the BD data obtained photographically. The good measurabilities in the case of these structures arise due to the large values of $\langle k \rangle$ for Co and I atoms with CuK α (see § 3.5 (f) for the values of $\langle k \rangle$).

The CCF of δ applicable to such a truncated data (denoted by $[N_{\delta}^c(\delta_0)]_{y_t}$) has been derived for a crystal of space group $P1$ by taking the Q -group to be ideally NC. The results for three cases, namely, $P=1$, MN and MC have been derived (Parthasarathy & Ponnuswamy, 1981). The CCF of δ for each case depends on the parameters k , σ_1^2 and y_t . $[N_{\delta}^c(\delta_0)]_{y_t}$ denotes the fractional number of reflections (in a given range of S) for which $\delta > \delta_0$ among those for which $y'_N \geq y_t$. However for discussing the effect of data truncation on the measurability the appropriate quantity is the fractional number of reflections for which $\delta > \delta_0$ and $y'_N \geq y_t$ relative to the population consisting of *all* the theoretically possible independent reflections in the given range of S . We shall denote this fraction by $f_{y_t}(\delta_0)$ and this is related to the CCF of δ for the truncated data by

$$f_{y_t}(\delta_0) = N_{y_N}^c(y_t) [N_{\delta}^c(\delta_0)]_{y_t}, \quad (22)$$

where $N_{y_N}^c(y_t)$ is the value of the CCF of y'_N at $y'_N = y_t$. The truncation limit y_t for the data of actual crystals would generally be in the neighbourhood of 0.2 (Ponnuswamy, 1979). The BD data for the observed reflections whose intensities are close to this truncation limit may not be very accurate. Hence in our discussion we shall assume that the BD data corresponding to reflections for which $y'_N > 0.3$ are sufficiently accurate to yield useful results. The curves of $f_{y_t}(0.1)$ vs σ_1^2 for $y_t = 0$ and 0.3 corresponding to typical values of k for the one-atom case are shown in Fig. 5. A study of these curves shows that data truncation due to unobserved reflections causes only a small decrease in measurability. For a typical situation in which k is small ($k=0.07$, say—this is close to the mean value of k for Cl with $\text{CuK}\alpha$, see Parthasarathi & Parthasarathy, 1974) and $P=1$ and $\sigma_1^2=0.3$ about 47% of the reflections are expected to have $\delta > 0.1$ when the

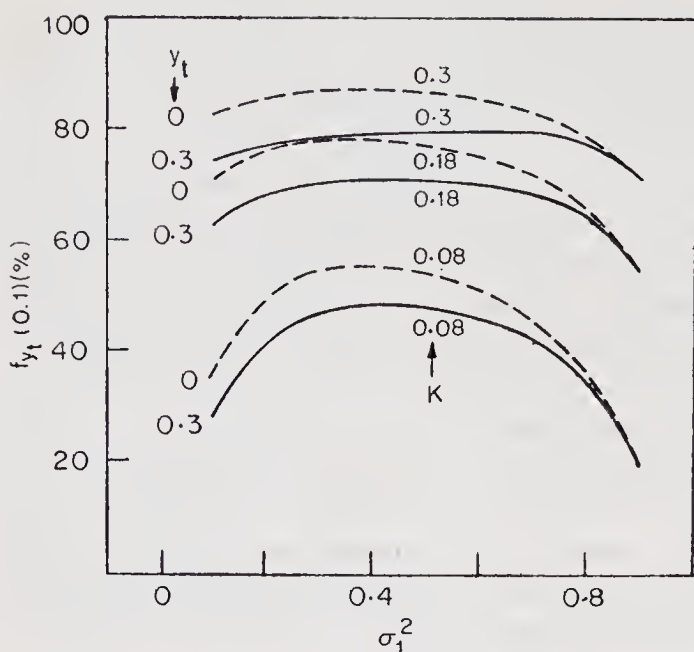


Fig. 5. $f_{y_t}(0.1) \times 100$ [i.e. the percentage of reflections for which $\delta > 0.1$ and $y'_N > y_t$] as a function of σ_1^2 for different fixed values of k corresponding to the untruncated data (i.e. $y_t = 0$) and to the data truncated at $y'_N = 0.3$. The broken lines are for the untruncated case and the solid lines are for the truncated case. The y -axis is shown displaced a little to the left. For the correct position of this axis, shift it to the right by parallel displacement such that it passes through the point $\sigma_1^2 = 0$.

data is truncated at $y_t = 0.3$ while this is 55% for the untruncated data. When k has a medium value ($k = 0.18$, say—this is a little less than the mean value of k for I with $\text{CuK}\alpha$) and $P=1$ and $\sigma_1^2=0.3$ these numbers for the truncated and untruncated data are 70% and 78% respectively. Thus though data truncation causes a small decrease in the measurability, it would not adversely affect the same.

(e) *Relevance of k over f''* . It is important to note that the ratio k but not the imaginary part f'' appears explicitly as a parameter of the CCF of X . Thus the measurability will be determined by the value of the ratio k but not by the absolute value of f'' alone. This is illustrated by two examples in Table 2. The results

Table 2. Examples showing the relevance of k over f'' for the measurability of BDs of a crystal

No.	P-atom	Q	$\langle \sigma_1^2 \rangle$	f''	$\langle k \rangle$	$f_{0.3}(0.1)$
1	Cl	50	24.1	0.702	8.2	45.8
	Br	50	62.3	1.283	6.5	37.2
2	Co	100	23.0	3.608	31.3	78.8
	I	100	67.0	6.835	22.2	74.1

Note: The values of $\langle \sigma_1^2 \rangle$, $\langle k \rangle$ and $f_{0.3}(0.1)$ are in percent. Values of f'' are taken from Srinivasan (1972). $\langle \sigma_1^2 \rangle$ and $\langle k \rangle$ are the average values for the range of $0 \leq \sin \theta / \lambda \leq 0.55 \text{ \AA}^{-1}$.

in Table 2 are for the one-atom case with $\text{CuK}\alpha$. The Q -atoms are chosen such that 80% of them are C, 10% are N and 10% are O. It is seen that though f'' for Br with $\text{CuK}\alpha$ is nearly twice that for Cl, the value of $\langle k \rangle$ for Br is slightly less than that for Cl. Thus in spite of f'' for Br being twice that for Cl the measurability for the Br compound is less than that for the Cl compound. Similar results are obtained for the second example in Table 2.

(f) *Choice of proper wavelength for atoms with $Z=10$ to 98 to optimize the measurability.* We have seen that the ratio k plays a prominent role in determining the measurability. We shall therefore study the manner in which k changes with atomic number Z of the atom. This incidentally helps us to understand the relative efficiency of two of the radiations generally used in crystal structure analysis, namely, $\text{CuK}\alpha$ and $\text{MoK}\alpha$ with respect to structures containing various types of heavy atoms. Since f_0 decreases while f' and f'' remain practically constant with increasing S , k will in general be an increasing function of S . For obtaining the theoretical CCF of any BD variable applicable to a given crystal the average value of k ($\langle k \rangle$, say) appropriate to the situation on hand must be used. Since f' and f'' depend on the wavelength of the radiation, $\langle k \rangle$ will also differ for the different wavelengths. Values of $\langle k \rangle$ for reflections upto $S=0.55 \text{ \AA}^{-1}$ (— this corresponds to a 2θ of about 115° for

$\text{CuK}\alpha$ and 45° for $\text{MoK}\alpha$) for atoms with $Z=10$ to 98 are shown in Fig. 6 for both $\text{CuK}\alpha$ and $\text{MoK}\alpha$ radiations. From Fig. 6 it is seen that $\text{CuK}\alpha$ is better suited than $\text{MoK}\alpha$ for all the elements except those for which Z ranges from 28 to 39 and 68 to 86. However for the elements in the range $Z=68$ to 86 the superiority of $\text{MoK}\alpha$ over $\text{CuK}\alpha$ is only marginal. For $\text{CuK}\alpha$ radiation Cr, Mn, Co, I, Ba, Sm, Gd, Pt, Au, Hg and Pb are some of the suitable elements and for $\text{MoK}\alpha$ Zn, Br, Rb, Pt, Au, Hg and Pb are suitable.

(g) *Measurability in macromolecular crystals.* The power of the X-ray anomalous scattering method for structure solution of macromolecular crystals (e.g. proteins) has been pointed out by Ramachandran & Parthasarathy (1965). We shall now discuss measurability in such complex structures. The values of $f_{0.3}$

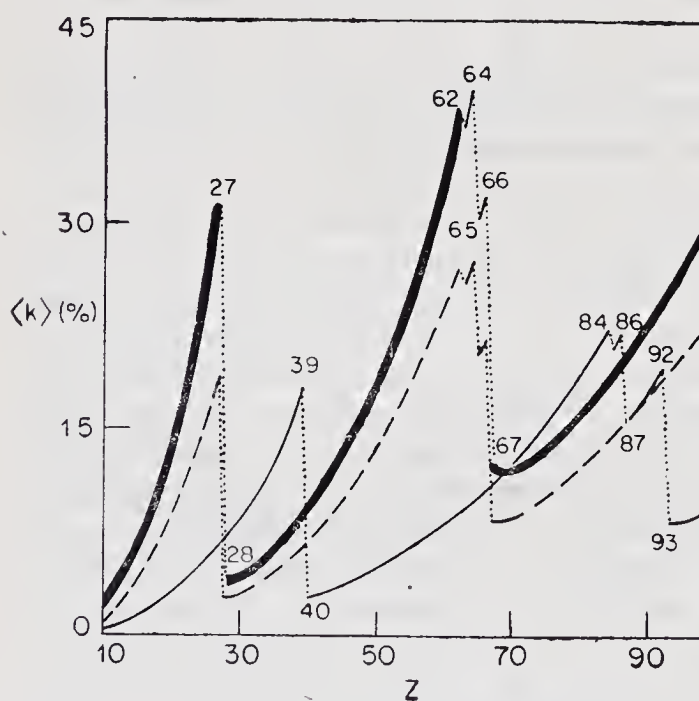


Fig. 6. Average value of k (appropriate to the range $0 \leq \sin\theta/\lambda \leq 0.55 \text{ \AA}^{-1}$) as a function of the atomic number Z for $\text{CuK}\alpha$ and $\text{MoK}\alpha$ radiations. The thick lines are for $\text{CuK}\alpha$ ($S_{\text{max}} = 0.55 \text{ \AA}^{-1}$), the thin lines are for $\text{MoK}\alpha$ ($S_{\text{max}} = 0.55 \text{ \AA}^{-1}$) and the broken lines are for $\text{CuK}\alpha$ ($S_{\text{max}} = 0.25 \text{ \AA}^{-1}$).

(0.1) are given in Table 3 for a number of typical situations. The results in Table 3 have been computed for crystals of space group $P1$ with one anomalous scatterer and Q light atoms per asymmetric unit. The Q -atoms are chosen such that 80% of them are carbons, 10% are nitrogens and 10% are oxygens. The values of $\langle \sigma_1^2 \rangle$ and $\langle k \rangle$ (for $\text{CuK}\alpha$) correspond to the data of 2\AA resolution. It is seen that for an asymmetric unit with one anomalous scatterer and 2000 light atoms the measurability values are 56% for Ba, 68% for Sm, 67% for Gd, 45% for Pt, 47% for Au, 49% for Hg, 53% for Pb and 67% for U. Thus with respect to measurability Sm, Gd, and U are better than Au, Pt, Hg and Pb and this mainly arises due to the larger values of $\langle k \rangle$ for the former elements. This shows that for realizing the full power of the anomalous scattering method the choice of proper wavelength is important. It is also seen from the examples in Table 3 that by exploiting the anomalous scattering in an optimum way (*i.e.* by choosing proper heavy atom derivative and by employing a proper wavelength for data collection) reasonably good measurability (more than 50%, say) can in general be obtained for proteins with even a few thousand atoms.

(h) *Measurability in crystals of moderate complexity.*

We shall now discuss the measurability in the case of crystals of moderate complexity (*i.e.* crystals containing a few hundred atoms per asymmetric unit) by taking a few examples. The values of $f_{0.3}$ (0.1) are given in Table 4 for a few typical cases and these correspond to the one-atom case and pertain to data for which $0 \leq S \leq 0.55\text{\AA}^{-1}$. It is seen from Table 4 that by using heavy atoms such as Fe, Co and I with $\text{CuK}\alpha$ one can obtain measurability as high as 70% even in structures with 500 atoms per asymmetric unit.

It is also seen from Table 4 that the measurability in structures containing about 40 atoms per asymmetric unit is nearly 40% with S as the anomalous scatterer and 47% with Cl (both with $\text{CuK}\alpha$). Thus

Table 3. *Measurability of Bijvoet differences to be expected in macromolecular crystals containing one anomalous scatterer and Q normal scatterers per asymmetric unit*

Atom			
$\langle k \rangle \%$	Q	$\langle \sigma_1^2 \rangle$	$f_{0.3} (0.1)$
Z		%	%
Ba	500	19.0	70.0
19.3	1000	10.5	65.3
56	1500	7.3	60.7
	2000	5.6	56.1
	2500	4.5	51.9
	3000	3.8	47.9
Sm	500	20.3	76.3
27.0	1000	11.3	73.4
62	1500	7.9	70.6
	2000	6.0	68.0
	2500	4.9	65.5
	3000	4.1	63.1
Gd	500	19.1	76.2
27.2	1000	10.6	73.0
64	1500	7.3	70.1
	2000	5.6	67.4
	2500	4.5	64.7
	3000	3.8	62.2
Pt	500	31.7	58.8
11.2	1000	18.9	54.9
78	1500	13.5	49.8
	2000	10.5	44.9
	2500	8.6	40.4
	3000	7.2	36.4
Au	500	32.5	59.9
11.6	1000	19.5	56.4
79	1500	13.9	51.8
	2000	10.8	47.2
	2500	8.9	42.8
	3000	7.5	39.0
Hg	500	33.1	60.9
12.0	1000	19.9	57.8
80	1500	14.3	53.6
	2000	11.1	49.2
	2500	9.1	45.0
	3000	7.7	41.3

Pb	500	34.2	63.2
13.0	1000	20.8	60.6
82	1500	14.9	57.1
	2000	11.6	53.4
	2500	9.5	49.7
	3000	8.1	46.2
U	500	39.0	71.5
18.5	1000	24.3	70.3
92	1500	17.7	68.6
	2000	13.9	66.9
	2500	11.5	65.0
	3000	9.7	63.2

Table 4. *Measurability of Bijvoet differences to be expected in crystals of small molecules and crystals of moderate complexity.*

Atom	Q	$\langle \sigma_1^2 \rangle$ %	$f_{0.3}$ (0.1) %
$ \langle k \rangle$ %			
P	20	39.3	34.9
5.6	30	30.2	32.7
	50	20.7	26.6
S	20	41.8	43.0
6.9	30	32.5	42.0
	50	22.4	37.3
	80	15.4	30.1
Cl	20	44.1	49.1
8.2	30	34.6	48.9
	50	24.1	45.8
	80	16.6	39.7
	100	13.8	35.8
Fe	50	39.3	77.6
26.5	100	24.5	76.7
	300	9.8	72.0
	500	6.1	67.7
Co	50	37.3	79.8
31.3	100	23.0	78.8
	300	9.0	74.5
	500	5.6	70.8

Br	20	80.4	26.2
6.5	30	73.3	31.9
	50	62.3	37.2
	80	51.0	40.0
	100	45.5	40.7
	200	29.6	38.7
I	50	80.1	71.1
22.2	100	67.0	74.1
	300	40.7	74.9
	500	29.3	74.3
Br	20	81.0	51.2
12.3	50	63.3	59.4
(MoK α)	100	46.5	61.6
	300	22.7	59.7
	500	15.0	55.2
Cu	30	65.3	43.4
7.9	50	53.2	46.7
(MoK α)	100	36.3	47.7
	200	22.2	43.1
	300	16.0	37.3

structures containing a *S*(or *Cl*) atom and 50 other light atoms per asymmetric unit can in principle be tackled by the anomalous scattering method.

(i) *Measurability in light atom structures.* We shall briefly discuss the measurability in light atom structures since this is important with respect to the determination of absolute configuration in such structures. The values of $f_y(\delta_o)$ corresponding to $\delta_o=0.03$, 0.05 and 0.1 and $y_t=0.2$ and 0.3 are given in Table 5 for the case $P=MN$ by taking the value of $\langle k \rangle$ to be 0.011 which corresponds to the mean value k for oxygen with CuK α radiation in the range $0 \leq S \leq 0.5 \text{ \AA}^{-1}$. The results in this table have been computed on the assumption that k of *C* for CuK α is negligible compared to that of *O* and these results may be applied to light atom structures containing *C* and *O*. A study of this table shows that a dozen reflections or more can in general be found for which $\delta > 0.05$ and $y'_N > 0.3$. Thus in spite of the data truncation due to

unobserved reflections BDs can be measured for a good number of reflections for establishing the absolute configuration (by the Bijvoet method) of an NC structure containing only light atoms.

(j) *Effect of space group symmetry.* In crystals with heavy-atoms the probability distribution of intensities depends on the presence of space-group symmetry elements other than the centre of symmetry (Karle & Hauptman, 1953; Hauptman & Karle, 1953). Foster & Hargreaves (1963*a, b*) have shown that space groups of the triclinic, monoclinic and orthorhombic systems can be classified into 7 categories* (called 1,2,3,...,7) based on the form of the trigonometric factors of the geometrical structure factor. Of these only the categories 1,3,5 and 6 belong to the NC case and hence here we need consider only these (see Foster & Hargreaves, 1963*b*). Though the probability distribution of X taking into account the space-group symmetry is not available, the expectation value of X has been calculated for crystals

Table 5. *Values of $f_{y_t}(\delta_0)$ (in %) as a function of σ_1^2 corresponding to $y_t = 0.2$ and 0.3 and $\delta_0 = 0.03, 0.05$ and 0.1 for the many-atom ($P=MN$) case when $k=0.011$.*

y_t	$\delta_0 \downarrow \sigma_1^2 \rightarrow$	0.1	0.2	0.3	0.4	0.5
0.2	0.03	5.8	10.8	13.9	15.7	16.2
	0.05	1.4	3.3	4.7	5.6	5.8
	0.10	0.0	0.3	0.5	0.7	0.7
0.3	0.03	4.2	8.3	11.1	12.8	13.3
	0.05	0.5	1.9	3.0	3.6	3.9
	0.10	0.0	0.0	0.1	0.2	0.2

*The results of this section can in principle be extended to space groups of higher symmetry by making use of the results of Wilson (1978), Shmueli & Wilson (1981) and Shmueli & Kaldor (1981) and this is in progress.

(of categories 1, 3, 5 and 6) containing 1 and 2 heavy atoms (all of one type) per asymmetric unit (Velmurugan & Parthasarathy, 1981). The values of $\langle X \rangle$ as a function of σ_1^2 for the cases $p=1$ and 2 corresponding to $k=0.1$ and 0.3 are shown in Fig. 7. The values of $\langle X \rangle$ as a function of k for $\sigma_1^2 = 0.1$ and 0.3 are shown in Fig. 8. From these figures we obtain the following results: (i) Among the very commonly occurring case of crystals containing one anomalous scatterer per asymmetric unit (*i.e.* $p=1$ case) in the region of general interest (*i.e.* $\sigma_1^2 < 0.7$) the triclinic space group $P1$ is the most favourable while the orthorhombic crystal of category 6 is the least favourable (the other conditions such as the complexity of the asymmetric unit, the type of heavy atom and the wavelength used being the same). The categories 3

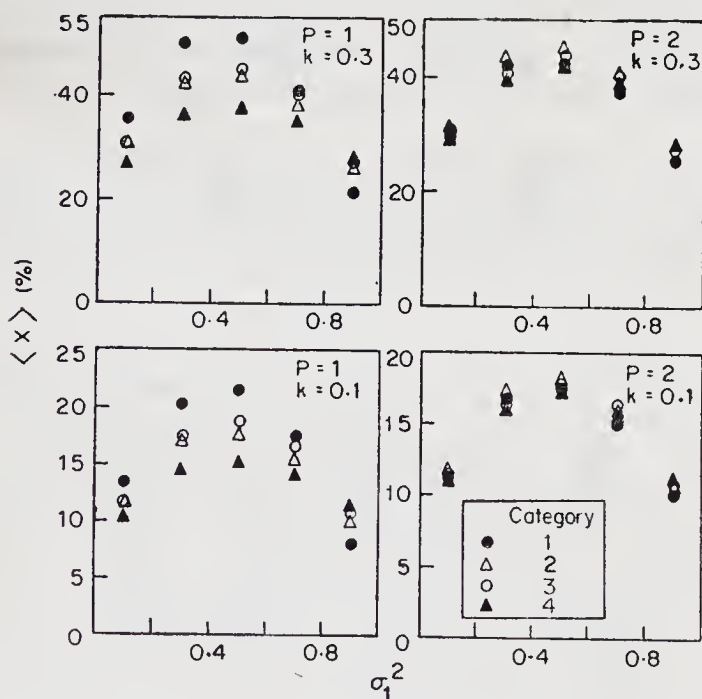


Fig. 7. Expectation value of X (in %) as a function of σ_1^2 (corresponding to $k = 0.1$ and 0.3) for the non-centrosymmetric space group categories 1, 3, 5 and 6 of the triclinic, monoclinic and orthorhombic systems containing $p (= 1$ or $2)$ anomalous scatterers per asymmetric unit.

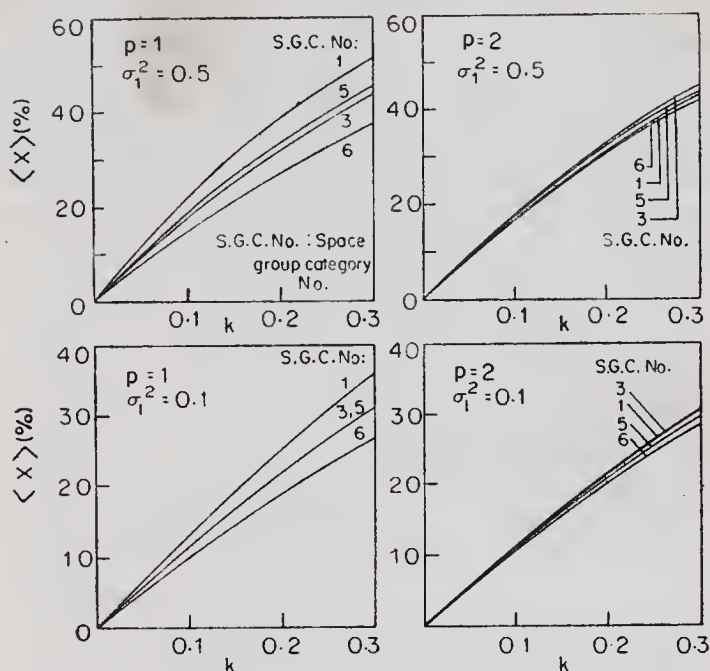


Fig. 8. Expectation value of X (in %) as a function of k (corresponding to $\sigma_1^2 = 0.1$ and 0.5) for the space group categories 1, 3, 5 and 6 containing p ($= 1$ or 2) anomalous scatterers per asymmetric unit.

and 5 are more or less equally effective and fall somewhat between the categories 1 and 6. (ii) For given k and σ_1^2 , as the number of anomalous scatterers per asymmetric unit increases from 1 to 2 the measurability decreases in category 1, increases somewhat in category 6 and remains practically unaffected in categories 3 and 5 and consequently the distinction between the various cases becomes less marked. For $p \geq 3$ the measurability is practically unaffected by space-group symmetry. These results are in agreement with those obtained by Parthasarathy & Ponnuswamy (1976) from a study of the expectation value of x . Incidentally it may be noted from these figures that the optimum conditions (i) and (ii) for measurability deduced for space-group $P1$ (see §3.5.(c)) are valid for space groups of higher symmetry.

3.6. *Measurability in the presence of degree of centrosymmetry*

So far we have considered the Q -group to be ideally NC in configuration. The Q -groups of actual crystals may exhibit different types of DCS. We shall discuss the measurability for the following three situations in crystals of space group $P1$ with one anomalous and Q similar normal scatterers in the unit cell: Situation (i): The Q -group consists of an ideally C part containing Q_c atoms besides an ideally NC part containing Q_n atoms (*i.e.* $Q = Q_c + Q_n$). We shall take the single P -atom to lie at a point which is significantly different from the centre of the Q_c group. Situation (ii): This is similar to situation (i) except that the single P -atom is now taken to lie exactly at the centre of the Q_c -group. For convenience we shall refer to the DCS met with in (i) as molecular DCS and that in (ii) as Type II DCS of the crystal (Parthasarathy & Parthasarathi, 1976*a*). Situation (iii): The Q -group consists of $Q/2$ atoms at \mathbf{r}_j ($j = 1$ to $Q/2$) and the other $Q/2$ atoms at $-\mathbf{r}_j + \Delta\mathbf{r}_j$ ($j = 1$ to $Q/2$), where $\Delta\mathbf{r}_j$ ($j = 1$ to $Q/2$) are $Q/2$ mutually independent 3-dimensional Gaussian vectors independent of the \mathbf{r}_j 's. The single P -atom is taken to lie at the centroid of the Q_c -group. We shall refer to this situation as Type I DCS of the crystal. We shall discuss the measurability for these three situations in §§3.6 (a), (b), & (c).

(a) *Molecular degree of centrosymmetry.* The DCS of the molecule (*i.e.* the Q -group) can be characterized by the quantity r defined by

$$r = \langle |F_{Qc}|^2 \rangle / \langle |F_Q|^2 \rangle,$$

which for a Q -group made up of similar atoms can be written as

$$r = Q_c / Q. \quad (23)$$

r is thus the ratio of the number of atoms in the C part of the Q -group to the total number of atoms in the Q -group. r is 0 when the Q -group is ideally NC and 1 when it is completely C and it takes intermediate

values corresponding to different DCS. The CCF of Δ depends on the parameters k , σ_1^2 and r (Parthasarathy & Parthasarathi, 1974;—see also Swaminathan & Srinivasan, 1974 for the limiting case $r = 1$). The curves of $N_{\Delta}^c(0.05)$ vs r are shown in Fig. 9 for different

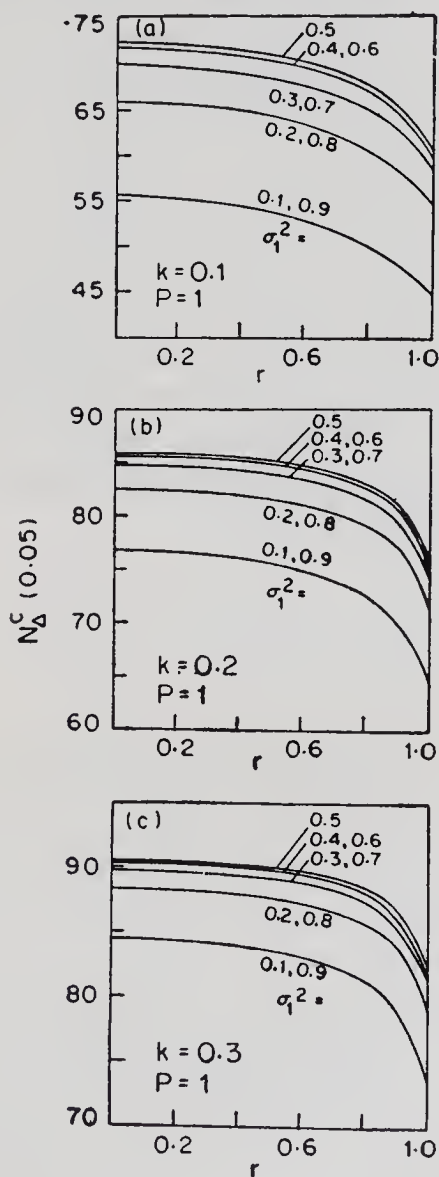


Fig. 9. $N_{\Delta}^c(0.05)$ (in %) as a function of r (i.e. the degree of centrosymmetry of the molecule) for different fixed values of σ_1^2 . (a), (b) and (c) correspond to $k = 0.1$, 0.2 and 0.3 respectively.

fixed values of σ_1^2 corresponding to $k=0.1, 0.2$ and 0.3 . These curves are nearly flat over a wide range of r ($r < 0.5$, say) showing thereby that the measurability would be practically unaffected even if 50% of the atoms in the molecule form a single C group. It is interesting to note that even if the whole molecule is completely C (now the combined P - and Q -groups have an overall NC configuration), in spite of the decrease in the value of $N_\Delta^c(0.05)$, there still exists enough percentage of reflections with measurable BD. For example, when $\sigma_1^2=0.4$ and $k=0.1$ we find $N_\Delta^c(0.05)$ to be 72% for $r=0$ and 60% for $r=1$. Thus in actual crystals the molecular DCS would pose no special problem with respect to measurability.

(b) *Crystal with type I degree of centrosymmetry.* A convenient measure of the Type I DCS of the crystal is $\langle |\Delta \mathbf{r}| \rangle_Q$. The CCF of Δ depends on the parameters k , σ_1^2 and D_Q (Parthasarathy & Parthasarathi, 1976b). where D_Q is defined as

$$D_Q = \langle \cos 2\pi \mathbf{H} \cdot \Delta \mathbf{r} \rangle_Q = \exp \left[-\frac{\pi^3}{4} H^2 \langle |\Delta \mathbf{r}|^2 \rangle_Q \right]. \quad (24)$$

The curves of $N_\Delta^c(0.05)$ vs $\langle |\Delta \mathbf{r}| \rangle_Q$ for the cases $k=0.1, 0.2$ and 0.3 (for $S=0.35 \text{ \AA}^{-1}$) are shown in Fig. 10. It is seen that when $\langle |\Delta \mathbf{r}| \rangle_Q=0.1 \text{ \AA}$, $k=0.1$, $S=0.35 \text{ \AA}^{-1}$ and $\sigma_1^2=0.5$ only 7% of the reflections have a $\Delta > 0.05$. Thus in crystals with a high Type I DCS [$\langle |\Delta \mathbf{r}| \rangle_Q < 0.1 \text{ \AA}$, say] the measurability is too low to be of use for structure determination. However, the breakdown of Friedel's Law can be detected if k is sufficiently large. This point is relevant to space-group determination using anomalous scattering in such crystals (Srinivasan & Vijayalakshmi, 1972).

(c) *Crystal with type II degree of centrosymmetry.* A convenient measure of the Type II DCS of the crystal is the quantity Q_c/Q ($=r$, say). The CCF of

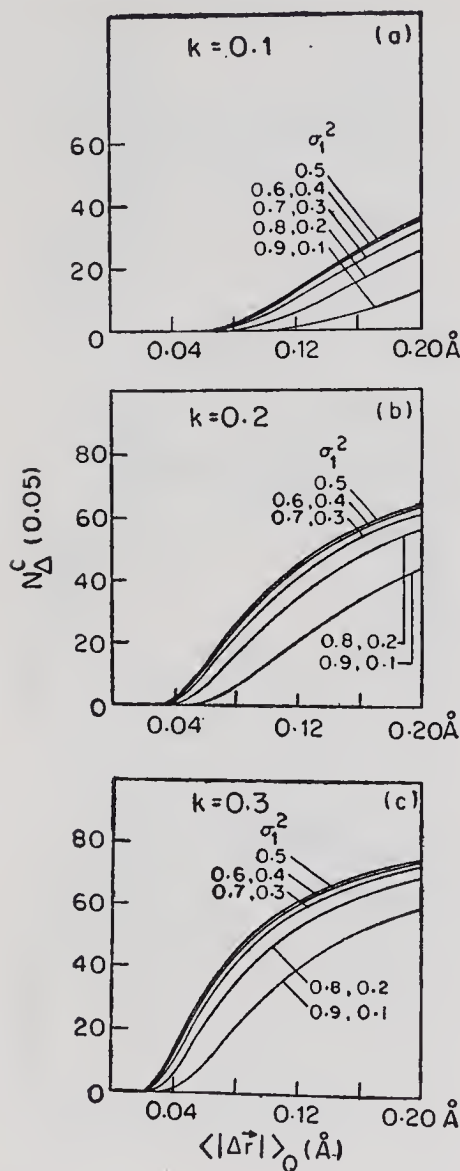


Fig. 10. $N_{\Delta}^c(0.05)$ (in %) as a function of $\langle |\Delta \vec{r}| \rangle_Q$ for different fixed values of σ_1^2 in crystals with Type I degree of centrosymmetry. (a), (b) and (c) correspond to $k = 0.1$, 0.2 and 0.3 respectively. These curves are for $\sin \theta/\lambda = 0.35 \text{ \AA}^{-1}$.

Δ depends on the parameters k , σ_1^2 and r (Parthasarathy & Parthasarathi, 1976b). Curves of $N_{\Delta}^c(0.05)$ vs r corresponding to $k = 0.1$, 0.2 and 0.3 for different σ_1^2 are shown in Fig. 11. It is seen that for medium and

C. S.—11

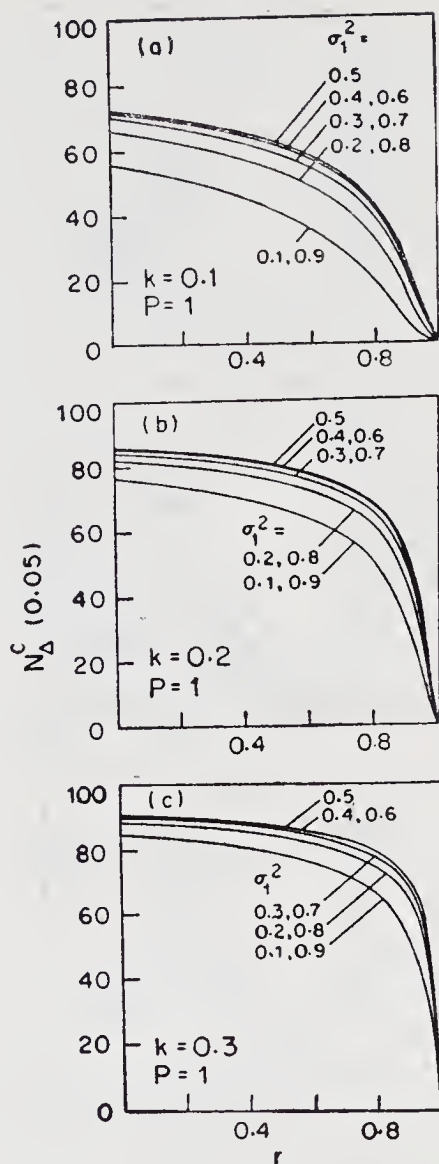


Fig. 11. $N_{\Delta}^c(0.05)$ (in %) as a function of r (i.e. the Type II degree of centrosymmetry of the crystal) for different fixed values of σ_1^2 . (a), (b) and (c) correspond to $k = 0.1$, 0.2 and 0.3 respectively.

large values of k ($k > 0.15$, say) the curves are practically flat for $r < 0.5$. However for small k (e.g. $k = 0.05$) the measurability decreases more or less systematically with increasing r . It is also seen that

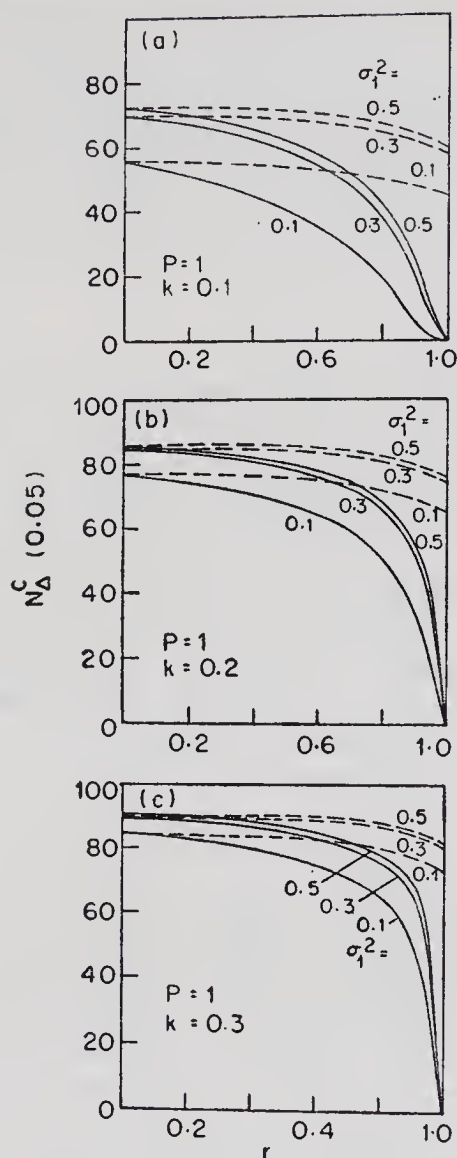


Fig. 12. Comparison of $N_{\Delta}^c(0.05)$ (in %) versus r for the two situations, namely, (i) the case of molecular DCS (dotted line) and (ii) crystal with Type II DCS (solid line). The numbers near the curves denote the values of σ_1^2 . (a), (b) and (c) correspond to $k = 0.1, 0.2$ and 0.3 respectively.

for $k=0.2$, $\sigma_1^2=0.5$ and $r=0.9$, $N_{\Delta}^c(0.05)$ is 55%. Thus crystals with Type II DCS as high as $r=0.9$ have enough percentage reflections with measurable BD for structure determination provided k is sufficiently large ($k > 0.15$, say).

A comparison of the results for a crystal with Type II DCS (§ 3.6(c)) and a crystal with molecular DCS (§3.6(a)) is interesting. From Fig. 12 it is seen that for given k and r the measurability in the former case is less than that in the latter. Further when r tends to 1 while the measurability in the former case is too low that in the latter case remains quite high even at $r=1$.

4. Measurability of Bijvoet difference of a reflection

The Bijvoet method (Bijvoet *et al.* 1951) of determining the absolute configuration in NC crystals consists of the following steps: (i) Completely determine and refine the crystal structure in a given configuration using the $h k l$ -data. (ii) Calculate the BR for all the reflections from the known atomic parameters and choose the top few (a dozen, say) reflections as optimum. (iii) Measure the BRs for these optimum reflections and compare these with the corresponding calculated values to arrive at the correct configuration. Thus it is necessary to keep the crystal until the structure is completely refined in order to collect the BR data for the few optimum reflections. We shall describe a statistical method which circumvents this necessity to some extent. This method is particularly useful when the anomalous scattering effect of the heavy atom compound is not pronounced* (*e.g.* organic molecules containing S, P or Cl). This method requires a knowledge of the intensity data for the $h k l$ -reflections and the magnitude $|F'_P|$ of the heavy atoms. Since the heavy atoms may be located from the Patterson, computed using the $h k l$ -intensity data, $|F'_P|$ may be taken to be known. Thus the present method can

*If the anomalous scattering effect is quite pronounced, enough number of reflections showing large BDs can be recognized even in the X-ray photographs. Hence the few reflections required can be chosen from a visual estimation of photographs during the stage of data collection.

be used to select the optimum reflections immediately after the heavy atoms have been located.

(a) *Definition of the measurability.* A probability measure, that is suitable for expressing the measurability of BD of a reflection for the present situation is the conditional CCF of X for given $|F'_N|$ and $|F'_P|$. That is

$$N_X^c(X_0; |F'_N|, |F'_P|) = P_r(X \geq X_0; |F'_N|, |F'_P|), \quad (25)$$

where X_0 is a particular value of X . The reflection for which the probability value $N_X^c(0.1; |F'_N|, |F'_P|)$ is high (greater than 0.75, say) may be chosen as suitable for BD measurement. Velmurugan, Parthasarathy & Parthasarathi (1979) have shown that

$$N_X^c(X_0; |F'_N|, |F'_P|) = 1 - \frac{2}{\pi I_0(\beta)} \times \int_0^{X_0} \frac{\cosh\left(\frac{\beta}{\alpha}(\alpha^2 - X^2)^{1/2}\right)}{(\alpha^2 - X^2)^{1/2}} dx = N_X^c(X_0; \alpha, \beta), \text{ say} \quad (26)$$

where α and β are defined to be

$$\alpha = \frac{4k |F'_N| |F'_P|}{|F'_N|^2 + k^2 |F'_P|^2}, \quad \beta = \frac{2 |F'_N| |F'_P|}{\langle |F_Q|^2 \rangle}. \quad (27)$$

Since α and β are single valued functions of $|F'_N|$ and $|F'_P|$ (26) can be taken to depend on the parameters α and β instead of $|F'_N|$ and $|F'_P|$. Values of k and $\langle |F_Q|^2 \rangle$ for a given reflection can be readily found from a knowledge of the unit cell content and unit cell parameters. Since the P -atoms are known, $|F'_P|$ will be known. When the anomalous scattering is not pronounced $|F'_N|$ may be equated to $F_{\text{obs}}(hkl)$. For a given X_0 (0.1, say) the probability value $N_X^c(0.1; \alpha, \beta)$ applicable to the reflection may be evaluated numerically from (26). The top few reflections having high probability values may be taken to be optimum for BD measurement.

A simple procedure for implementing (26) can be devised by studying the properties of the equation

$$N_X^c(X_0; \alpha, \beta) = p. \quad (28)$$

For given values of $X_0 (=0.1, \text{ say})$ and p , (28) represents a curve in the α, β -plane. The curves corresponding to $p=0.75, 0.8, \dots$ are shown in Fig. 13 for the case $X=0.1$. For given X_0 and p , the curve starts from a point $(\alpha_{\min}, 0)$ on the α -axis and increases thereafter as α increases. Thus, for given X_0 and p , relation (28) is not possible if $\alpha < \alpha_{\min}$. This means that given the values of X_0 and p the reflections for which $\alpha < \alpha_{\min}$ cannot show a $BR > X_0$ with a probability value p or more. This property may be used to eliminate reflections which are not suited for BD measurement. For any α ($> \alpha_{\min}$) there is a unique β ($=\beta_t$, say) for which (28) is satisfied. This β_t can be obtained by solving (28) numerically.

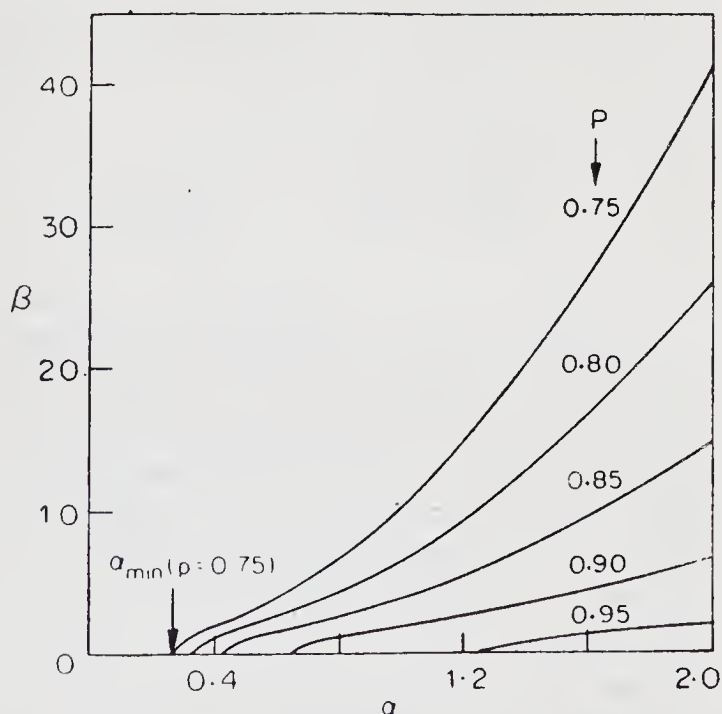


Fig. 13. Functional relationship between α and β satisfying the equation $N_X^c(0.1; \alpha, \beta) = p$ corresponding to $p = 0.75, 0.8, \dots, 0.95$.

(b) *Implementation of the theoretical results.* The following procedure may be used to select the few optimum reflections: (i) Locate the heavy atoms (from the Patterson, say) using the intensity data of hkl -reflections. (ii) Calculate the value of α from (27) for each reflection using the known values of $F_{\text{obs}}(hkl)$ and $|F'_p|$. Reject the reflections for which $\alpha < 0.26$ i.e. for which $\text{Pr}(X \geq 0.1; \alpha, \beta) < 0.75$. Let N_1 be the number of remaining reflections. (iii) Find the values of β_t for these N_1 reflections from their values of α by linear interpolation from the results in Table 6. Calculate the values of β from (27) for these N_1 reflections. Reject the reflections for which $\beta > \beta_t$. Let N_2 be the remaining number of reflections. (iv) Calculate the probability values $N_X^c(0.1; \alpha, \beta)$ for these N_2 reflections by bilinear interpolation using the results in Table 7. Order these N_2 reflections with decreasing probability values and choose the top dozen reflections as optimum for BD measurement.

The results of application of (26) in the case of a few actual crystals are shown in Table 8. The probability values $N_X^c(0.1; \alpha, \beta)$ are calculated for all the reflections directly from (26). The reflections are then arranged in the decreasing order of probability. The top ten reflections with the highest probability values are given. The last column contains the corresponding

Table 6. β as a function of α satisfying the equation $N_X^c(0.1; \alpha, \beta) = 0.75$

α	β	α	β	α	β	α	β
0.26	0.00	0.70	5.24	1.15	13.72	1.60	26.33
0.30	0.97	0.75	5.98	1.20	14.91	1.65	27.98
0.35	1.40	0.80	6.77	1.25	16.16	1.70	29.69
0.40	1.86	0.85	7.61	1.30	17.46	1.75	31.45
0.45	2.32	0.90	8.50	1.35	18.81	1.80	33.26
0.50	2.82	0.95	9.44	1.40	20.21	1.85	35.12
0.55	3.35	1.00	10.43	1.45	21.66	1.90	37.03
0.60	3.93	1.05	11.47	1.50	23.17	1.95	39.00
0.65	4.56	1.10	12.57	1.55	24.72	2.00	41.01

Table 7. *Conditional complementary cumulative function $N_X^c(0.1; \alpha, \beta) \times 1000$ as a function of α and β*

$\beta \searrow \alpha \rightarrow$ \downarrow	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0
0	667	784	839	872	893	909	920	929	936	942	947	951	954	958	960	963	965	966	968
1	607	740	806	845	871	889	903	914	922	929	935	940	945	948	951	954	957	959	961
2	494	656	740	791	826	850	869	883	895	905	913	919	925	930	934	938	942	945	947
3	395	578	678	741	783	814	837	855	869	881	891	899	906	913	918	923	927	931	934
4	319	515	627	698	747	782	809	830	847	861	872	882	890	898	904	910	915	919	923
5	261	462	584	662	716	756	785	809	828	843	856	867	877	885	892	898	904	909	913
6	216	418	546	630	689	732	764	790	811	828	842	854	864	873	881	888	894	900	905
7	180	380	513	602	664	710	745	773	795	813	829	842	853	863	871	879	885	891	897
8	150	347	483	576	642	690	728	757	781	800	817	830	842	853	862	870	877	884	889
9	126	318	456	552	621	672	711	742	767	788	805	820	833	844	853	862	870	876	882
10	106	291	432	530	602	655	696	728	755	776	795	810	823	835	845	854	862	869	876
11	90	268	409	510	584	639	681	715	743	766	785	801	815	827	838	847	855	863	870
12	76	247	388	491	567	624	668	703	732	755	775	792	807	819	830	840	849	857	864
13	65	227	368	473	551	609	655	691	721	745	766	783	799	812	823	834	843	851	858
14	55	210	350	456	535	595	642	680	711	736	757	775	791	805	817	827	837	845	853
15	47	194	333	440	521	582	631	669	701	727	749	768	784	798	810	821	831	840	848
16	40	180	317	425	507	570	619	659	691	718	741	760	777	791	804	815	825	835	843

17	34	166	302	411	494	558	608	649	682	710	733	753	770	785	798	810	820	829	838
18	29	154	288	397	481	546	598	639	673	701	725	746	763	779	792	804	815	824	833
19	25	143	275	384	469	535	588	630	665	694	718	739	757	773	787	799	810	820	829
20	21	133	263	372	457	525	578	621	656	686	711	732	751	767	781	794	805	815	824
21	18	124	251	360	446	514	568	612	648	679	704	726	745	761	776	789	800	810	820
22	16	115	240	349	435	504	559	604	641	671	697	720	739	756	771	784	796	806	816
23	14	107	229	338	425	495	550	596	633	664	691	714	733	750	766	779	791	802	811
24	12	99	219	327	415	485	542	588	626	657	684	708	728	745	761	774	787	798	807
25	10	93	210	317	405	476	533	580	618	651	678	702	722	740	756	770	782	793	804
26	9	86	201	308	396	467	525	572	611	644	672	696	717	735	751	765	778	789	800
27	7	80	192	299	387	459	517	565	605	638	666	691	712	730	746	761	774	785	796
28	6	75	184	290	378	451	509	558	598	632	660	685	707	725	742	757	770	782	792
29	6	70	177	281	370	443	502	551	591	626	655	680	702	721	737	752	766	778	789
30	5	65	169	273	362	435	495	544	585	620	649	675	697	716	733	748	762	774	785
31	4	61	162	265	354	427	487	537	579	614	644	669	692	711	729	744	758	770	782
32	4	57	156	257	346	420	480	531	573	608	638	664	687	707	725	740	754	767	778

Table 8. Top ten reflections with the highest probability values N_X^c ($0.1; |F'_N|, |F'_P|$) and the corresponding observed values for the Bijvoet ratio for some actual crystal.

<i>L-lephedrine hydrochloride</i>				<i>L-tyrosine hydrochloride</i>				<i>L-lysine hydrochloride dihydrate</i>			
<i>h</i>	<i>k</i>	<i>l</i>	N_X^c (0.1) X_{obs}	<i>h</i>	<i>k</i>	<i>l</i>	N_X^c (0.1) X_{obs}	<i>h</i>	<i>k</i>	<i>l</i>	N_X^c (0.1) X_{obs}
3	4	0	90.3% 0.44	4	4	0	88.9% 0.41	1	2	0	93.9% 0.42
3	2	0	89.8 0.54	1	9	0	81.8 0.59	2	8	0*	91.4 0.07
7	3	0	89.0 0.12	5	4	0	78.8 0.19	4	6	0	90.4 0.36
5	2	0	82.5 0.19	2	9	0	78.4 0.20	3	4	0	90.3 0.42
9	3	0*	81.9 0.05	5	6	0	78.4 0.21	4	8	0*	89.7 0.00
6	3	0	78.9 0.27	10	6	0	78.4 0.13	5	6	0	87.4 0.13
3	5	0	78.7 0.24	8	5	0*	78.1 0.07	1	4	0*	87.2 0.06
7	1	0*	78.4 0.01	2	5	0	77.3 0.17	5	8	0	86.5 0.15
7	4	0	77.3 0.54	5	2	0	75.1 0.14	1	12	0	80.5 0.21
11	3	0*	77.1 0.09	9	3	0	74.1 0.34	1	8	0*	78.4 0.02
								0	11	1	71.8 0.12

Note: The reflection for which the prediction is wrong is shown with an asterisk. $N_X^c(0.1) = N_X^c(0.1; |F'_N|, |F'_P|)$.

observed values of the BR. For each crystal the reflection for which the prediction is wrong is shown with an asterisk (*). The present method is seen to be successful in 73% of cases on the average.

The result in (26) can also be used to select the reflections for BD measurement for the purpose of structure determination. For details one may refer to the original paper.

5. Conclusions

The main results on the measurability of BDs of a crystal may be summarized as follows: (a) The optimum conditions for the measurability of BDs of a crystal are: (i) k should be as large as possible and (ii) σ_1^2 should be in the neighbourhood of 0.5. (b) Measurability depends on the ratio k rather than on f'' . (c) Measurability is influenced more by k than by σ_1^2 . A proper choice of wavelength is therefore of great importance to realize the full power of anomalous scattering. (d) Data truncation due to unobserved reflections causes only a small decrease in the measurability. Hence even in complex non-centrosymmetric structures containing a few thousand atoms measurability of the order of 50% or more may be obtained by exploiting the anomalous scattering phenomenon in an optimum way. (e) In crystals with one anomalous scatterer per asymmetric unit the triclinic category 1 is the most favourable while the orthorhombic category 6 is the least favourable (the other conditions such as the complexity of the asymmetric unit, the type of heavy atom and the wavelength used being the same). The distinction between the space-group categories becomes lesser and gets evened out as the number of anomalous scatterers per asymmetric unit increases. (f) The measurability is practically unaffected even if 50% of the atoms in a molecule form a single centrosymmetric group. (g) In crystals with a high Type I

degree of centrosymmetry the measurability is too low to be of use for structure determination.

The author wishes to express his thanks to Professor A. J. C. Wilson for the invitation to deliver the talk on the measurability of BDs at a Microsymposium held during the XII Congress and General Assembly of the International Union of Crystallography, Ottawa. He is thankful to Professor K. Venkatesan and Dr. K. K. Chacko for useful discussion. His thanks are due to Dr. M. N. Ponnuswamy and Mr. D. Velmurugan for help in the preparation of the manuscript.

References

- BIJVOET, J. M. (1952). *Computing Methods and Phase Problem in X-ray Crystal Analysis* edited by R. PEPINSKY
- BIJVOET, J. M. (1954). *Nature* **173**, 888.
- BIJVOET, J. M. (1955). *Endeavour*. **14**, 71.
- BIJVOET, J. M., PEERDEMAN, A. F. & VAN BOMMEL, A. J. (1951). *Nature* **168**, 271.
- DALE, D., HODGKIN, D. C. & VENKATESAN, K. (1963). *Crystallography and Crystal Perfection* edited by G. N. RAMACHANDRAN, London: Academic Press.
- FOSTER, F. & HARGREAVES, A. (1963a). *Acta Cryst.* **16**, 1124.
- FOSTER, F. & HARGREAVES, A. (1963b). *Acta Cryst.* **16**, 1133.
- HALL, S. R. & MASLEN, E. N. (1965). *Acta Cryst.* **18**, 265.
- HAUPTMAN, H. & KARLE, J. (1953). *Acta Cryst.* **6**, 136.
- JAMES, R. W. (1958). *The Optical Principles of the Diffraction of X-rays*. London: G. Ball.
- KARLE, J. & HAUPTMAN, H. (1953). *Acta Cryst.* **6**, 131.
- PARTHASARATHY, S. (1967). *Acta Cryst.* **22**, 98.
- PARTHASARATHY, S. & PARTHASARATHI, V. (1973). *Acta Cryst.* **A29**, 428.
- PARTHASARATHI, V. & PARTHASARATHY, S. (1974). *Acta Cryst.* **B30**, 1375.
- PARTHASARATHY, S. & PARTHASARATHI, V. (1974). *Acta Cryst.* **A30**, 649.
- PARTHASARATHY, S. & PARTHASARATHI, V. (1976a). *Acta Cryst.* **A32**, 57.
- PARTHASARATHY, S. & PARTHASARATHI, V. (1976b). *Acta Cryst.* **A32**, 768.

- PARTHASARATHY, S. & PONNUSWAMY, M. N. (1976). *Acta Cryst.* **A32**, 302.
- PARTHASARATHY, S. & PONNUSWAMY, M. N. (1981). *Acta Cryst.* **A37**, 153.
- PARTHASARATHY, S., RAMACHANDRAN, G. N. & SRINIVASAN, R. (1964). *Curr. Sci.* **33**, 637.
- PARTHASARATHY, S. & SRINIVASAN, R. (1964). *Acta Cryst.* **17**, 1400.
- PEERDEMAN, A. F. & BIJVOET, J. M. (1956). *Acta Cryst.* **9**, 1012.
- PONNUSWAMY, M. N. (1979). Ph.D. Thesis, University of Madras.
- RAMACHANDRAN, G. N. & PARTHASARATHY, S. (1965). *Science* **150**, 212.
- RAMACHANDRAN, G. N. & RAMAN, S. (1956). *Curr. Sci.* **25**, 348.
- RAMACHANDRAN, G. N. & SRINIVASAN, R. (1970). *Fourier Methods in Crystallography*, New York: John Wiley.
- RAMASESHAN, S. (1964). *Advanced Methods in Crystallography*, edited by G. N. RAMACHANDRAN, London: Academic Press.
- RAMASESHAN, S. & ABRAHAMS, S.C. (Editors)(1975). *Anomalous Scattering*, Munksgaard.
- SHMUELI, U. & KALDOR, U. (1981). *Acta Cryst.* **A37**, 76.
- SHMUELI, U. & WILSON, A. J. C. (1981). *Acta Cryst.* **A37**, 342.
- SRINIVASAN, R. (1972). *Advances in Structure Research by Diffraction Methods*. Vol. IV, edited by W. HOPPE & R. MASON, West Germany: Pergamon Press.
- SRINIVASAN, R. & VIJAYALAKSHMI, B. K. (1972). *Acta Cryst.* **B28**, 2615.
- SIM, G. A. (1964). *Acta Cryst.* **17**, 1072.
- SWAMINATHAN, P. & SRINIVASAN, R. (1974). *Acta Cryst.* **A30**, 702.
- VELMURUGAN, D., PARTHASARATHY, S. & PARTHASARATHI, V. (1979). *Acta Cryst.* **A35**, 463.
- VELMURUGAN, D. & PARTHASARATHY, S. (1981) (unpublished).
- WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318.
- WILSON, A. J. C. (1978). *Acta Cryst.* **A34**, 986.
- YOW-LAM OH & MASLEN, E. N. (1966). *Acta Cryst.* **20**, 852.
- ZACHARIASEN, W. H. (1965). *Acta Cryst.* **18**, 714.

Intensity Statistics and Non-Independence

BY A. J. C. WILSON

*Department of Physics, University of Birmingham,
Birmingham B15 2TT, England*

Abstract

The usual expressions for the probability distribution of X-ray reflexions are derived on the assumption that the contributions of individual atoms to the structure factor are independent. In reality the finite size and stereochemical properties of atoms prevent complete independence. Central-limit theorems still apply, but are there valid expansions of the Gram-Charlier or Edgeworth type?

Statistics and non-independence

While the statisticians are still here I should like to raise a question about the incorporation of stereochemical considerations into the mathematical expressions for the probability distributions of X-ray reflexions when the number of atoms is too small for the central-limit argument (Wilson, 1949) to be applied. If the contributions of individual atoms to the expressions for the structure factor,

$$F_{hkl} = \sum_{j=1}^n f_j \exp \{2\pi i (hx_j + ky_j + lz_j)\}, \quad (1)$$

are assumed to be independent, then the probability distribution can be expanded as a series involving orthogonal polynomials; a paper by Shmueli & Wilson (1981) may be consulted for a detailed discussion and references. In reality, of course, the finite sizes and the stereochemical properties of atoms

prevent complete independence. There are central-limit theorems valid for 'almost independent' variables (Bernstein, 1922) and for variables dependent only on finite number $f(n)$ of their neighbours (Bernstein, 1927). This second case seems plausible for the crystallographic application, as the positions of neighbouring atoms are closely correlated, but most molecules have enough flexibility to prevent appreciable correlation between widely separated atoms. Harker (1953) suggested that protein molecules could be regarded as consisting of a number of 'globs', with atomic positions within a 'glob' being correlated, but without correlation between 'globs'. This seems analogous to $f(n)$ dependence.

If the generalized central-limit theorem is applied to the case of non-independent atomic contributions, the functional forms of the distributions deduced by Wilson (1949) are retained but the distribution parameter is increased from

$$\Sigma = \sum_{j=1}^n |f_j|^2 \quad (2)$$

(Wilson, 1942) to $\langle I \rangle$, the actual local average of the intensity of reflexion. This was tacitly assumed by many authors, and French & Wilson (1978) made it an explicit empirical postulate; their postulate for the value of the parameter has been justified theoretically (Wilson, 1981). In the independent case the distribution function for a finite number of atoms can be expressed as the sum of the ideal distribution (the central limit) and a series of correction terms, each correction term being the product of three factors:

- (i) the ideal distribution;
- (ii) one of a set of orthogonal polynomials; and
- (iii) a function of the moments of the distribution.

In general such series, considered as functions of the number n of terms summed, are only asymptotic, but

intensity distributions appear to fulfil the conditions for genuine convergence. My question is: Are there analogous series for the sum of n non-independent random variables? One might guess that factors (i) and (ii) would remain, but the functions of the moments in (iii) would be altered in a manner analogous to the substitution of $\langle I \rangle$ for Σ as the distribution parameter in the central-limit expression.

References

- BERNSTEIN, S. (1922). *Math. Ann.* **85**, 237–241.
BERNSTEIN, S. (1927). *Math. Ann.* **97**, 1–59.
FRENCH, S. & WILSON, K. (1978). *Acta Cryst.* **A34**, 517–525.
HARKER, D. (1953). *Acta Cryst.* **6**, 731–736.
SHMUELI, U. & WILSON, A. J. C. (1981). *Acta Cryst.* **A37**, 342–353.
WILSON, A. J. C. (1942). *Nature*, **150**, 151, 152.
WILSON, A. J. C. (1949). *Acta Cryst.* **2**, 318–321.
WILSON, A. J. C. (1981). *Acta Cryst.* **A37**, 808–810.

Statistics of Recorded Counts

BY J. L. DE BOER

*Laboratory for Chemical Physics, University of
Groningen, Nijenborgh 16, 9747 AG Groningen,
The Netherlands*

Introduction

In many publications the distribution functions of intensities influenced by random (counting) errors have been discussed and estimates of standard deviations in measured intensities have been worked out (Wilson, 1978, 1980, and references therein). For the experimentalist in X-ray diffraction, this field of research is like an utopia, almost unrealistic as he can aim for but hardly reach it by his experimental data. The present paper will show that especially strong reflections find obstacles on their way. In addition, we will see that for weak reflections, background can be a real bottleneck.

As an example accurate measurements for electron-density studies will be discussed. One of the objects of this field of research is to map deformation properties, such as the deformation density

$$\Delta\rho(\mathbf{r}) = \frac{1}{V} \sum \{ F_0(\mathbf{H}) - F_c(\mathbf{H}) \} \exp(-2\pi i \mathbf{H} \cdot \mathbf{r}), \quad (1)$$

or the deformation potential

$$\Delta\phi(\mathbf{r}) = \frac{-1}{4\pi V} \sum [\{ F_0(\mathbf{H}) - F_c(\mathbf{H}) \} \exp(-2\pi i \mathbf{H} \cdot \mathbf{r})] / (\sin \theta / \lambda)^2. \quad (2)$$

The formulae are for a centrosymmetric crystal; $F_c(\mathbf{H})$ is based on a reference model of spherically symmetric atoms, which we assume to be known. The measurements have been set up with care. For the present discussion we consider a data set in which 'intrinsic'

systematic errors in the intensities, like extinction, absorption, TDS and multiple diffraction have been eliminated. For the diffractometer work the following conditions were fulfilled after careful checks: homogeneous monochromatic X-ray beam, uniform response of counter surface, correct alignment of diffractometer, counter width and scan range chosen such that full integrated intensities could be measured. The question arises whether under these carefully chosen measuring conditions the errors in the intensities are due to counting statistics alone. With the assumption that the answer will be 'yes' standard deviations for very strong and very weak reflections will be estimated on the basis of approximate formulae which, however, are sufficient to obtain an insight in the essential features of these quantities.

Estimates of standard deviations

The intensity $I(\mathbf{H}; \text{net})$ of a reflection \mathbf{H} is given by

$$I(\mathbf{H}; \text{net}) = [k/t(\mathbf{H})] [I_p(\mathbf{H}) - m(\mathbf{H}) I_b(\mathbf{H})]; \quad (3)$$

k is a scale factor which is equal for all reflections and which obeys the relation $k \sim I_0^{-1} v^{-1}$ where I_0 is the intensity of the primary beam and v is the volume of the crystal; $t(\mathbf{H})$ = time spent on the 'peak' of reflection \mathbf{H} ; $m(\mathbf{H}) = t(\mathbf{H}; \text{peak})/t(\mathbf{H}; \text{background})$. $I_p(\mathbf{H})$ and $I_b(\mathbf{H})$ are the number of counts measured for peak and background respectively. The variance $\sigma^2 [I(\mathbf{H}, \text{net})]$ is given by

$$\sigma^2 [I(\mathbf{H}; \text{net})] = [k^2/t^2(\mathbf{H})] [I_p(\mathbf{H}) + m^2(\mathbf{H}) I_b(\mathbf{H})]. \quad (4)$$

The $|F_0(\mathbf{H})|^2$ value

$$|F_0(\mathbf{H})|^2 = (Lp) I(\mathbf{H}; \text{net}), \quad (5)$$

has the variance

$$\sigma^2 [|F_0(\mathbf{H})|^2] = [k^2(Lp)^2/t^2(\mathbf{H})] [I_p(\mathbf{H}) + m^2(\mathbf{H}) I_b(\mathbf{H})]. \quad (6)$$

(a) *Very strong reflections*

These reflections obey the assumption $m(\mathbf{H}) I_b(\mathbf{H}) \ll I_p(\mathbf{H})$ so that

$$I(\mathbf{H}; \text{net}) \simeq k I_p(\mathbf{H})/t(\mathbf{H}), \quad (7)$$

and

$$\sigma^2 [|F_0(\mathbf{H})|^2] = [k(\text{Lp})^2/t(\mathbf{H})] I(\mathbf{H}; \text{net}). \quad (8)$$

With $\sigma(|F|) = \sigma(|F|^2)/2|F|$ we obtain from (8) and (5)

$$\sigma(|F_0(\mathbf{H})|) = k^{\frac{1}{2}}(\text{Lp})^{\frac{1}{2}}/2t^{\frac{1}{2}}(\mathbf{H}) = k^{\frac{1}{2}}/2t_n^{\frac{1}{2}}(\mathbf{H}), \quad (9)$$

with $t_n(\mathbf{H}) = t(\mathbf{H})/(\text{Lp})$

Conclusion: For this category of reflections $\sigma(|F_0(\mathbf{H})|)$ does not depend on $I(\mathbf{H}; \text{net})$!

(b) *Very weak reflections*

Rees (1976) has shown that for reflections with $|F_0(\mathbf{H})|^2 \leq \sigma[|F_0(\mathbf{H})|^2]$ a good estimate for the standard deviation is given by

$$\sigma[|F_0(\mathbf{H})|] = \sigma^{\frac{1}{2}}[|F_0(\mathbf{H})|^2]/2. \quad (10)$$

For not too small backgrounds, the near-zero reflections obey the approximation

$$I_p(\mathbf{H}) \simeq m(\mathbf{H}) I_b(\mathbf{H}). \quad (11)$$

Combination of (6), (10) and (11) gives

$$\sigma(|F_0(\mathbf{H})|) = k^{\frac{1}{2}} |F_b(\mathbf{H})|^{\frac{1}{2}} [1 + m(\mathbf{H})]^{\frac{1}{2}}/2t_n^{\frac{1}{2}}(\mathbf{H}), \quad (12)$$

in which $|F_b(\mathbf{H})|^2$ is the number of background counts per second corrected for Lp.

Conclusion: For this category of reflections it is the background which determines the standard deviation!

The necessity of reducing the background

The accuracy required for the individual reflections depends on the property which one wants to study.

For the potential, the accuracy must increase with $[(\sin \theta)/\lambda]^{-2}$. For the density, on the other hand, all standard deviations should be the same. Getting the weak reflections measured with adequate accuracy can be a real problem as (under the condition where (12) is valid) increase in measuring time is very inefficient to improve accuracy. In contrast to the very strong reflections (formula 9) where it costs counting 4 times longer to improve σ by a factor of 2, the present situation requires counting 16 times longer! A much more efficient way is to reduce $|F_b|$, which can be done by monochromatisation and/or collimation of the incoming and outgoing beam. Table 1 illustrates the reduction in background which can be achieved in this way.

In the limiting case with $|F_b(\mathbf{H})|^2 \rightarrow 0$, when approximation (11) no longer holds, $\sigma(|F_0(\mathbf{H})|)$ can be obtained directly from (6) and (10) with (3) and (5):

$$\sigma(|F_0(\mathbf{H})|) = [k/t_n(\mathbf{H})]^{1/4} |F_0(\mathbf{H})|^{1/2}/2. \quad (13)$$

In this case, for equal standard deviations, the measuring times $t_n(\mathbf{H})$ are proportional to $|F_0(\mathbf{H})|^2$ and become now very short for the very weak reflections! Realisation of the limit $F_b \rightarrow 0$ is also the preferable way to solve the problems arising for reflections with negative $I(\mathbf{H}; \text{net})$ intensities, the omission of which gives a bias in the maps.

Is counting statistics the only error source?

In a recent experiment on forsterite, Mg_2SiO_4 , we wanted to measure the low-order reflections with extreme accuracy to study the deformation potential $\Delta\phi(\mathbf{r})$.

The conditions were such that for the strong reflection 020 an accuracy of 0.04% was required. The diffractometer was programmed accordingly, but

Table 1. *Illustration of background reduction attained on a CAD-4F diffractometer, for measurements with Mo radiation. Any sequence of three numbers a/b/c gives the observed counts for background left, peak and background right respectively in a θ -2 θ scan.*

A. Reduction effects of monochromator.		
Apart from substitution of the beta filter by the monochromator (plus flatmaker, see Helmholtz & Vos, 1977), all other measuring circumstances are equal. Data are given for a reflection <i>nh</i> , <i>nk</i> , <i>nl</i> and a 'general' reflection <i>hkl</i> of H-propylmorpholinium (TCNQ) 2.		
Beta filter, without monochr.	refl. 0 0 8	refl. $\bar{1}$ 0 8
With monochr.	2119/11386/944	136/10905/586
	87/7583/106	94/11912/171
B. Reduction effects of collimators.		
Data are given for a low-and a high-order reflection of forsterite, Mg ₂ SiO ₄ , measured with monochromatised Mo radiation. <i>N.B.</i> It is seen that the smallest collimator nozzle (no. 3) is in fact a bit too small for the crystal used (7% loss of net intensity)		
Without nozzle and detector coll.	refl. 0 2 1	refl. 0 6 2
Coll. nozzle no. 1; no extra detector coll.	16993/959457/11047	4114/1067489/12763
Coll. nozzle no. 2; no extra detector coll.	8691/941045/ 5872	3451/1070122/12266
Coll. nozzle no. 3; no extra detector coll.	5531/924871/ 3907	3390/1060882/11738
Coll. nozzle no. 3; <i>plus</i> extra detector coll.	1902/848049/ 1319	2849/1004506/ 9781
	610/843509/ 477	2144/ 997286/ 8690

repeated measurements of the reflection revealed an r.m.s. variation of ten times larger. To discover the reason for this variation the reflections 020, 021, 062, 133 and 211 were measured alternately. In comparison with the other 4 reflections, 020 showed a strong variation from -1.2 to 5.6% (for discussion see below). Results for the other 4 reflections are given in Fig. 1. With use of these reflections as intensity reference reflections, corrections according to the full line could have been made. We see, however, that apart from the 'jump' in intensity there are pseudo-random variations which are clearly larger than expected on the basis of the 0.1% statistical error aimed for in the on-line experiment. In experiments which would have aimed at a statistical accuracy of, say, 1% , may be not even the 'jump' and certainly not the additional pseudo-random variations would have been noticed. Nevertheless they would have been there, making the distribution function for the reflections different from

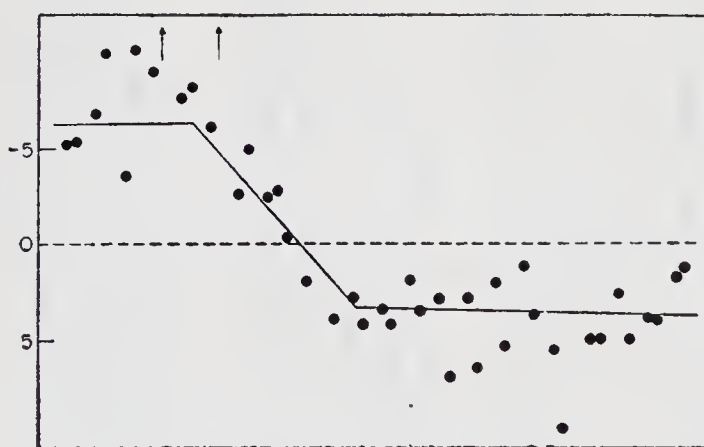


Fig. 1. Relative intensities of 4 reflections measured consecutively in a certain period of time. Sequence \equiv time axis stands horizontal (dashed). Vertical is $1000 [I/\langle I \rangle - 1]$. Each of the individual measurements has a standard deviation of $\approx 0.1\%$. The figure shows a 'jump' within 2 hours of about 1% and moreover a spread around the full line larger than expected.

the theoretical distribution function based on counting statistics.

Analysis of step-scan values for the two extreme values of reflection 0 2 0 (Fig. 2) showed that this discrepancy is mainly due to a non-constant motor speed. It may be possible to remove part of it by apparatus reconstruction. But even then, it cannot be excluded that diffractometers are subject to small variations in their electronical and mechanical parts, giving intensities which do not obey distribution functions based on counting statistics. Whether or not the deviations are important, depends on the accuracy required for the experiment.

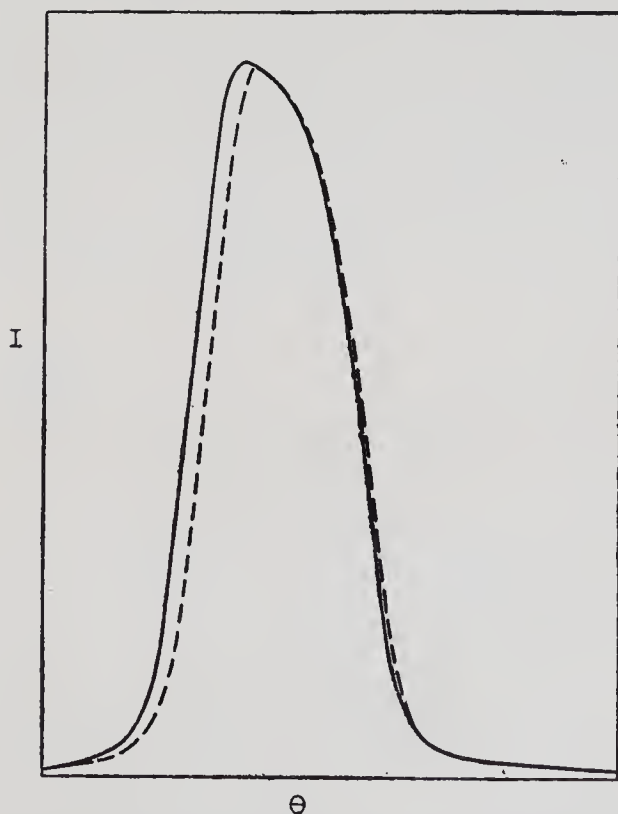


Fig. 2. Profiles (θ - 2θ scan) of the two extreme situations of reflection 0 2 0. There is some difference in height but the main difference is found in width (on an absolute scale however being small, not more than 0.015 degree discrepancy).

References

- HELMHOLDT, R. B. & VOS, A. (1977). *Acta Cryst.* **A33**, 456–465.
REES, B. (1976). *Acta Cryst.* **A32**, 483–488.
WILSON, A. J. C. (1978). *Acta Cryst.* **A34**, 474–475.
WILSON, A. J. C. (1980). *Acta Cryst.* **A36**, 929–936.

Alternatives to R Tests

BY STUART M. ROTHSTEIN

*Department of Chemistry, Brock University,
St. Catharines, Ontario, Canada L2S 3A1*

Abstract

Three alternatives to R -tests are compared in a computer simulation study of power and robustness: Rothstein, Richardson, and Bell's jackknife test on the R -factor ratio, Arvesen's jackknife test for the correlation coefficient, and Pitman's test for the correlation coefficient which uses Pearson's statistic. It is found that unless one could improve the approximate null-distributions for Arvesen's and Pitman's test, Rothstein *et al.*'s procedure is best, having simulated probabilities of Type I error closest to the test's nominal α and being reasonably robust and powerful, for all distributions considered.

Introduction

Hamilton's test on the R -factor ratio,

$$\mathcal{R} = R_{\text{I}}/R_{\text{II}}, \quad (1)$$

where R_{I} and R_{II} are the residuals associated with models I and II, is well-known to crystallographers as a means of determining if there is a significant difference in the R factors, and hence the goodness of fit of the models (Hamilton, 1964, 1965).

Hamilton was able to derive the null-distribution of \mathcal{R} by assuming that the structure factors are linear in the parameters (2), and the hypothesis under test is a linear relationship on the parameters (3).

$$\mathbf{F} = \mathbf{A}\mathbf{x} + \epsilon, \quad (2)$$

$$\mathbf{Q}\mathbf{x} = \mathbf{Z}, \quad (3)$$

where $F = (|F_i|_0 - |F_i|_c)$, $|F_i|_0$ and $|F_i|_c$ are the observed and calculated values of the i th structure factor, $A = (\partial |F_i|_c / \partial x_j)$, $x = (\Delta x_j)$, x_j is the j th parameter and Δx_j is the correction to be applied to the j th parameter. The vector $\epsilon(2)$ represents a collection of random, non-systematic discrepancies between the structure factor differences and the model values.

The null distribution of \mathcal{R} is given by

$$\mathcal{R} \sim (b \mathcal{F}_{b, N-m, \alpha} / (N - m) + 1)^{1/2}, \quad (4)$$

where b is the rank of Q (3), N is the number of observations, m is the number of parameters in the least constrained model, and \mathcal{F} is the F -statistic critical value at the α -level of significance. Derivation of this result requires $\epsilon(2)$ to be independent and identically normally distributed.

Our objective in this paper is to report various alternatives to Hamilton's test, that is, discriminate between Models I and II, where we need not assume linear models (2), or a linear hypothesis under test (3), and where minimal assumptions are made concerning the distribution of $\epsilon(2)$.

Procedures

Rothstein, Richardson & Bell's (1978) [RRB's] jackknife test is one such alternative, already applied to crystallographic problems. The test is based on an underlying model dealing with the distribution of Δ_i for each of Models I and II, where

$$\Delta_i^I \equiv \sqrt{\omega_i} (|F_i|_0 - |F_i|_c^I), \quad (5)$$

and ω_i is the weight associated with the i th observation; an analogous expression defines Δ_i^{II} . In particular, $[\Delta_i^I, \Delta_i^{II}]$ are assumed bivariate distributed with variance covariance matrix

$$\Sigma = \begin{bmatrix} \sigma_I^2 & \sigma_{I, II} \\ \sigma_{I, II} & \sigma_{II}^2 \end{bmatrix}, \quad (6)$$

and $\sigma_{I, II} \neq 0$.

The null hypothesis that the two models are equally good descriptions of the observed data

$$H_0: \sigma_I^2 = \sigma_{II}^2, \quad (7)$$

is tested *versus* the alternative that Model II is the better model

$$H_a: \sigma_I^2 > \sigma_{II}^2. \quad (8)$$

Rejecting H_0 is evidence for rejecting Model I. RRB's procedure involves computing

$$\mathcal{L}_i = N \ln (\mathcal{R}^2) - (N-1) \ln (\mathcal{R}_{-i}^2), \quad (9)$$

where \mathcal{R}_{-i}^2 means to compute $\mathcal{R}^2(1)$ where the i th reflection has been deleted from the calculation of \mathcal{R}^2 . Their test statistic is given by

$$Q' = \overline{\mathcal{L}} / V', \quad (10)$$

where $\overline{\mathcal{L}}$ is the average value of \mathcal{L}_{-i} (9) and

$$V'^2 \equiv \sum (\mathcal{L}_{-i} - \overline{\mathcal{L}})^2 / [N(N-1)]. \quad (11)$$

RRB conjectured that under H_0 (7), Q' is approximately t -distributed with $N-1$ degrees of freedom, *i.e.* asymptotically normal. This paper, in part, tests this conjecture in computer simulation studies.

Another viable alternative to Hamilton's test concerns the statistical correlation of linear combinations of Δ^I and Δ^{II} :

$$A_i \equiv \Delta_i^I + \Delta_i^{II}, \quad (12)$$

$$B_i \equiv \Delta_i^I - \Delta_i^{II}. \quad (13)$$

H_0 (7) and H_a (8) become equivalent to

$$H'_0: \rho_{AB} = 0, \quad (14)$$

$$H'_a: \rho_{AB} > 0, \quad (15)$$

respectively, where ρ_{AB} is the correlation coefficient for A and B .

Arvesen's (1969) jackknife test for the correlation

coefficient is believed to be effective in testing the hypothesis $\rho=0$, based on the results of an empirical study using 1000 samples ($N=25,50$) published by Johnson (1978).

For the sake of comparison, Pitman's (1939) test of the hypothesis H'_0 (14) will also be considered. This test uses Pearson's correlation coefficient r and is a 'normal theory' procedure; the distribution of A_i and B_i must be bivariate normal. Accordingly, it violates a major criterion for a viable alternative to Hamilton's test, that of minimal assumptions being made concerning the distribution of ϵ (2).

Computer simulation

We will soon publish (Bell, Rothstein, & Li, 1982) an assessment of the performance of RRB's test of H_0 (7) and Arvesen's test of H'_0 (14) by a Monte Carlo simulation involving 20,000 experiments, each generating two samples ($N=12$) of pseudo random numbers (Δ^I and Δ^{II} values) drawn from the bivariate normal, bivariate (1% and 5%) Cauchy-contaminated normal, and bivariate gamma distributions. As Bell *et al.* are publishing relevant technical details of the simulation, only typical results, sufficient to generalize the relative performance of the tests will be cited here.

When the null hypothesis is true, a test performs well if the simulated significance levels α (probabilities of a Type I error) are close to the test's nominal α . Bell *et al.*'s results (Table 1) show that RRB's jackknife test gives significance levels which are closer to the theoretical values than those obtained from Arvesen's procedure for each of the three symmetric distributions, and although the errors are comparable for the skewed distribution (bivariate gamma), RRB's test is more conservative. It is clear that RRB stay consistently closer to the nominal α than do Arvesen's or Pitman's tests, and the latter two tests consistently reject H_0 (or H'_0) too often when the hypothesis is true.

Table 1. Empirical type I error probabilities using 20,000 samples ($N=12$).

$\sigma_X^2 = \sigma_Y^2 = 1.0$ and $\sigma_{XY} = 0.9$					
	RRB ^a	Arvesen ^b			Pitman ^c
	0.05	0.01	0.10	0.05	0.01
0.10	0.05	0.01	0.10	0.05	0.01
<i>Normal distribution</i>					
0.098	0.052	0.014	0.125	0.076	0.030
<i>Normal—1% Cauchy contamination</i>					
0.102	0.057	0.017	0.128	0.079	0.030
<i>Normal—5% Cauchy contamination</i>					
0.116	0.057	0.014	0.150	0.093	0.033
<i>Gamma distribution</i>					
0.076	0.033	0.008	0.121	0.063	0.018
			0.130	0.086	0.039

^aRothstein *et al* (1978); ^bArvesen (1969); ^cPitman (1939).

Table 2. Empirical power using 20,000 samples ($N=12$) with $\sigma_X^2=1.2$, $\sigma_Y^2=1.0$, $\sigma_{XY}=0.9$

	RRB ^a	Arvesen ^b			Pitman ^c		
0.10	0.05	0.01	0.10	0.05	0.01	0.10	0.01
<i>Normal distribution</i>							
0.213	0.121	0.029	0.213	0.121	0.029	0.226	0.034
<i>Normal—1% Cauchy contamination</i>							
0.204	0.108	0.023	0.204	0.117	0.031	0.210	0.024
<i>Normal—5% Cauchy contamination</i>							
0.191	0.108	0.026	0.200	0.115	0.029	0.185	0.015
<i>Gamma distribution</i>							
0.209	0.130	0.035	0.198	0.125	0.039	0.212	0.019

^aRothstein, *et al* (1978); ^bArvesen (1969); ^cPitman (1939).

With the null hypothesis false, a test performs well if the empirical power, $1-\beta$, is large. (β is the empirical probability of accepting the false hypothesis.) After adjusting the critical values for each test to make the simulated α agree with the nominal α , the empirical power obtained by Bell *et al.* appears in Table 2. Not surprisingly, Pitman's test performs best for the bivariate normal and mildly (1%) Cauchy contaminated normal distributions, otherwise it has consistently less power. For the other distributions, the power of Arvesen's test is consistently better than RRB's, although only in the third decimal place.

In conclusion, if one could improve the null distributions for Arvesen's test and Pitman's test, the former would be the best for the non-normal distributions considered by Bell *et al.* and the latter best for normally distributed data. In the absence of such results, however, RRB's test is the best procedure, consistently performing well under H_0 and being reasonably robust and powerful.

I wish to emphasize that the work described in this paper is drawn from collaborations and/or discussion with Mr W. D. Bell, Ms H. L. Gordon, and Drs W. -K. Li, M. F. Richardson, and D. M. Thompson. Our research was supported, in part, by grants from the Natural Sciences and Engineering Research Council of Canada.

References

- ARVESEN, J. N. (1969). *Ann. Math. Stat.* **40**, 2076-2100.
 BELL, W. D., ROTHSTEIN, S. M. & LI, W.-K. (1982). *J. stat. comput. simul.*, to be published.
 HAMILTON, W. C. (1964). *Statistics in Physical Science. Estimation, Hypothesis Testing, and Least Squares*. New York: Roland Press.
 HAMILTON, W. C. (1965). *Acta Cryst.* **18**, 502-519.
 JOHNSON, N. J. (1978). *J. Am. Stat. Assoc.* **73**, 536-544.
 PITMAN, E. J. G. (1939). *Biometrika* **31**, 9-12.
 ROTHSTEIN, S. M., RICHARDSON, M. F. & BELL, W. D. (1978). *Acta Cryst.* **A34**, 969-974.

The Residual Function R_2 as Discriminator Criterion in Structure Determination

BY A. T. H. LENSTRA

*Department of Chemistry, University of Antwerp
(U.I.A.), Universiteitsplein 1, B-2610 Wilrijk, Belgium*

Introduction

In automated procedures to determine a crystal structure, one obviously needs criteria to discriminate between correct and incorrect structure models. Such structure models, to be checked on their reliability, can be obtained in many ways. Heavy-atom analysis, rotation and translation searches or a direct-method routine are the most likely sources for tentative structure models. The structures to be tested can be organic as well as inorganic. Therefore the discriminator function must be very flexible in terms of applicability in experimental situations. As a consequence mathematical discriminator functions must be preferred above *e.g.* chemical criteria, of which the applicability is restricted to one class of chemical compounds, say organic structures. From our own experience (Lenstra 1973; van de Mierop 1979) we know that the heavy-atom analysis can be successfully automated handling the residual R_2 as a mathematical indicator function. R_2 is defined as

$$R_2 = \sum_H (I_N - I_n)^2 / \sum_H I_N^2 \quad (1)$$

where the observed structure contains N atoms and the tentative model contains n ($n \leq N$) atoms. In general the model to be tested has g atoms at correct atomic sites and f atoms at incorrect positions. This situation will be denoted as (g, f) with $g + f = n$.

The functional behaviour of R_2 has been evaluated in many papers (Wilson 1969, 1974, 1976; Lenstra

1973, 1974, 1979; Parthasarathy & Parthasarathi 1972; Srinivasan & Parthasarathy 1976; Van de Microop & Lenstra 1978; Petit, Lenstra & Van Loock 1981; Petit & Lenstra 1981). For theoretical convenience the crystal structure is regarded as a system of non-vibrating point atoms. This has the advantage that the scattering power of the atoms is independent of the Bragg angle θ . Due to this (1) can be rephrased giving

$$R_2 = \sum_H (E_N^2 - \sigma_1^2 E_n^2)^2 / \sum_H E_N^4, \quad (2)$$

where
$$\sigma_1^2 = \sum^n f_j^2 / \sum^N f_j^2$$

In the rest of this paper we discuss some properties of R_2 or related functions. The parameters, which are especially dealt with in terms of their influence on R_2 , are: the size of the structure model ($\sigma_1^2 \sim n/N$) and the threshold a . This threshold is applied on E_N^2 -data only, and its introduction means that all observed intensities with $E_N^2 < a$ are omitted in the practical computation of R_2 . The main reason to introduce the parameter a is to reduce the time necessary to enumerate R_2 , which makes R_2 better applicable in practical work.

In §§ 2 and 3, R_2 is calculated as a simple point estimator. The logical line goes analogous to all references cited. Within this framework we will study:

- (i) situation $(n, 0)$ versus $(0, n)$. From a mathematical point of view this is the simplest situation. For practical use this description is only of interest for the traditional rotation-search procedures in reciprocal space.
- (ii) situations (g, f) . This is the general description apt to any model. It will be shown that $R_2(g, f)$ is a predictable quantity.

For a proper use of R_2 in applied crystallography it is evident that the knowledge of a point estimator is insufficient but better than nothing. So in § 4 $P(R_2)$ is discussed indirectly. It is shown that any moment $\langle (R_2 - R_2)^a \rangle$ can be calculated. For brevity only the space group $P1$ is dealt with.

In §5 rotation search is discussed, as an example to show the validity of the present theory.

2. Correct and incorrect models with n atoms ($n \leq N$)

Replacing the summations in (2) by the number of observations times the actual averages, we find

$$R_2 = \langle (E_N^2 - \sigma_1^2 E_n^2)^2 \rangle / \langle E_N^4 \rangle. \quad (3)$$

The angular brackets indicate an average overall experimental intensities available. A reduction in the number of reflections used to calculate actual R_2 values speeds up the application procedure. To accelerate the reliability test itself we follow the standard practice implemented in rotation search (*e.g.* Tollin & Rossmann, 1966), *i.e.* eliminate all E_N^2 values below a certain threshold a . Then $R_2(a)$ can be written as

$$R_2(a) = (\langle E_N^4 \rangle_a - 2\sigma_1^2 \langle E_N^2 E_n^2 \rangle_a + \sigma_1^4 \langle E_n^4 \rangle_a) / \langle E_N^4 \rangle_a. \quad (4)$$

R_2 is, of course, only a proper mathematical indicator function if its final numerical value remains predictable, though now as a function of the threshold a . To obtain numerical values for $R_2(a)$ the actual averages in (4) are replaced by distribution averages. To avoid unwanted complexity due to the overwhelming number of available intensity distributions, we have to restrict ourselves. In this paper we tackle the problems in terms of equal-atom structures in the space groups $P1$ and $P\bar{1}$. The basic distributions which

we use throughout this chapter are the Wilson distributions:

$$P(E) = 2 E \exp[-E^2] \quad \text{for } P1$$

and

$$P(E) = \left(\frac{2}{\pi}\right)^{1/2} \exp\left[-\frac{E^2}{2}\right] \quad \text{for } P\bar{1}.$$

For the sake of completeness it should be noted that these functions are only strictly valid for structures with a large number of randomly located atoms in the unit cell. Consequently, these distributions are asymptotic approximations of the more precise functions, given by Srinivasan & Parthasarathy (1976), in which the number of atoms determines the more exact algebraic formulation.

2.1. *Incorrect models (0, n); unrelated case*

In this situation E_N and E_n are independent of each other. So the joint probability distribution $P(E_N, E_n)$ is simply the product $P(E_N)P(E_n)$. For unrelated model (4) can be simplified to

$$R_2(a) = (\langle E_N^4 \rangle_a + \sigma_1^4 \langle E_n^4 \rangle - 2\sigma_1^2 \langle E_N^2 \rangle_a \langle E_N^4 \rangle_a) \quad (5)$$

The necessary distribution averages are easily obtained, because

$$\langle E^q \rangle_a = \int_{\sqrt{a}}^{\infty} E^q P(E) dE / \int_{\sqrt{a}}^{\infty} P(E) dE.$$

The individual intensity moments are summarised in Table 1. Substitution of these moments in (5) gives for the space group $P1$

$$R_2(a) = [a^2 + 2a(1 - \sigma_1^2) + 2(\sigma_1^4 - \sigma_1^2 + 1)] \star [a^2 + 2a + 2]^{-1} \quad (6)$$

Table 1. Relevant moments for $P1$ and $P\bar{1}$ in function of the threshold “ a ” for the two extreme situations $(n, 0)$ and $(0, n)$. σ_1^2 is given by $n|N$ and

$$Q = \sqrt{\frac{2a}{\pi}} \exp(-a/2) / \operatorname{erfc}\left(\sqrt{\frac{a}{2}}\right). \quad n \text{ and } N \text{ are supposed to be large.}$$

Space group	$P1$		$P\bar{1}$	
Model	$(n, 0)$	$(0, n)$	$(n, 0)$	$(0, n)$
$\langle E_n^2 \rangle_a$	$1 + \sigma_1^2 a$	1	$1 + \sigma_1^2 Q$	1
$\langle E_n^4 \rangle_a$	$\sigma_1^4 a^3 + 2\sigma_1^2 (2 - \sigma_1^2) a + 2$	2	$3 + \sigma_1^2 (6 + \sigma_1^2 (a - 3)) Q$	3
$\langle E_N^2 E_n^2 \rangle_a$	$\sigma_1^2 a^3 + (1 + \sigma_1^2) a + 1 + \sigma_1^2$	$1 + a$	$1 + 2\sigma_1^2 + (1 + \sigma_1^2 (2 + a)) Q$	$1 + Q$
$\langle E_N^2 \rangle_a$	$1 + a$		$1 + Q$	
$\langle E_N^4 \rangle_a$	$a^3 + 2a + 2$		$3 + (3 + a) Q$	

For the space group $P\bar{1}$ we find

$$R_2(a) = 1 + \left[-2 \sigma_1^2 \operatorname{erfc} \sqrt{\frac{a}{2}} - 2 \sigma_1^2 \sqrt{\frac{2a}{\pi}} \right. \\ \left. \exp(-a/2) + 3 \sigma_1^4 \operatorname{erfc} \sqrt{\frac{a}{2}} \right] \star \left[3 \operatorname{erfc} \sqrt{\frac{a}{2}} \right. \\ \left. + (3+a) \sqrt{\frac{2a}{\pi}} \exp(-a/2) \right]^{-1}. \quad (7)$$

The functional behaviour of $R_2(a)$ is visualized in Figs. 1 ($P1$) and 2 ($P\bar{1}$). It is clear that for a given model of size σ_1^2 , R_2 depends strongly on the applied threshold.

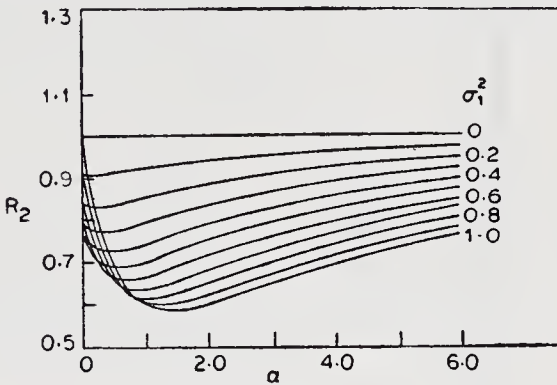


Fig. 1. R_2 for the unrelated case in space group $P1$.

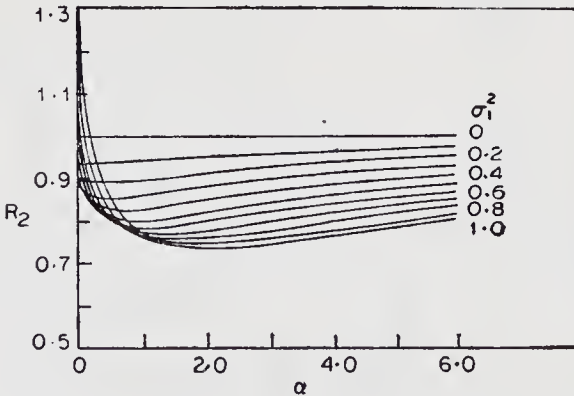


Fig. 2. R_2 for the unrelated case in space group $P\bar{1}$.

2.2. Correct models ($n, 0$); related case

For correct structure models the structure-factor equation is

$$E_N = E_n + E_r,$$

where E_r represents the unknown part of the structure (size: $N-n$). Since E_N and E_n are now mutually dependent the joint distribution $P(E_N, E_n)$ is no longer the simple product of $P(E_N)$ and $P(E_n)$. The joint distributions we need are (see Srinivasan & Parthasarathy, 1976):

$$P(E_N, E_n) = \frac{4E_N E_n}{\sigma_2^2} \exp\left[-\frac{E_N^2 + E_n^2}{\sigma_2^2}\right] I_0\left(\frac{2\sigma_1 E_N E_n}{\sigma_2^2}\right) \dots, \quad (8)$$

and

$$P(E_N, E_n) = \frac{2}{\pi\sigma_2} \exp\left[-\frac{E_N^2 + E_n^2}{\sigma_2^2}\right] \cosh\left(\frac{\sigma_1 E_N E_n}{\sigma_2^2}\right), \quad (9)$$

which are derived for the space group $P1$ and $P\bar{1}$ respectively.

I_0 is a modified Bessel function of the first kind of order zero and $\sigma_2^2 = 1 - \sigma_1^2$. The intensity moments necessary to evaluate R_2 from (4) are listed in Table 1.

The algebraic derivation of these moments is illustrated with one single example, viz.

$$\langle E_N^2 E_n^2 \rangle_a = \frac{A}{B} = \frac{\int_{\sqrt{a}}^{\infty} \int_0^{\infty} E_N^2 E_n^2 P(E_N, E_n) dE_n dE_N}{\int_{\sqrt{a}}^{\infty} \int_0^{\infty} P(E_N, E_n) dE_n dE_N}. \quad (10)$$

(a) *The space group $P1$* : Substituting (8) in (10) and rewriting the Bessel function as its series (Abramovitz & Stegun, 1968), we have

$$A = \frac{4}{\sigma_2^2} \sum_{k=0}^{\infty} \frac{\sigma_1^{2k}}{\sigma_2^{4k} (k!)^2} \cdot \int_0^{\infty} E_n^{2k+3} \exp\left[-\frac{E_n^2}{\sigma_2^2}\right] \star$$

$$\int_{\sqrt{a}}^{\infty} E_N^{2k+3} \exp\left[-\frac{E_N^2}{\sigma_2^2}\right] dE_N dE_n.$$

Substituting $Z_N = E_N^2$ and $Z_n = E_n^2$ and using the following standard integrals

$$\int_a^\infty x^m e^{-nx} dx = e^{-na} \sum_{r=0}^m \frac{m! a^{m-r}}{(m-r)! n^{r+1}}$$

$$\int_0^\infty x^n e^{-bx} dx = \frac{n!}{b^{n+1}} \quad \text{for } b > 0 \text{ and } n \text{ positive,}$$

the resulting double series $\sum_k \sum_r$ can be shown to be equal to:

$$e^{-a} [\sigma_1^2 (a^2 + a + 1) + 1 + a].$$

The normalisation factor B is given by e^{-a} , a result easily found using the marginal Wilson distribution $P(E_N)$.

Consequently: $\langle E_N^2 E_n^2 \rangle_a = [\sigma_1^2 (a^2 + a + 1) + 1 + a]$ as given in Table 1.

(b) *The Space group $P\bar{1}$* : The recipe goes as follows. Substituting (9) in (10) and applying the definition of $\cosh x = (e^x + e^{-x})/2$, the numerator A is given by

$$A = \frac{\sigma_2^5}{\pi} \int_{\sqrt{a}/\sigma_2}^\infty u^2 \exp\left[-\frac{u^2}{2}\right] \int_0^\infty v^2 \exp\left[-\frac{v^2}{2}\right] \star$$

$$[\exp(\sigma_1 uv) + \exp(-\sigma_1 uv)] dv du,$$

$$\text{where } u = \frac{E_N}{\sigma_2} \text{ and } v = \frac{E_n}{\sigma_2}.$$

Applying the standard integrals

$$\int_0^\infty \exp[-(at^2 + bt + c)] dt = \frac{1}{2} \sqrt{\frac{\pi}{a}}$$

$$\exp\left[\frac{b^2 - ac}{a}\right] \operatorname{erfc}\left(\frac{b}{\sqrt{a}}\right),$$

$$i^n \operatorname{erfc}(z) = \frac{2}{\sqrt{\pi}} \int_z^\infty \frac{(t-z)^n}{n!} e^{-t^2} dt.$$

$$\text{So: } A = (1 + 2 \sigma_1^2) \operatorname{erfc} \sqrt{\frac{a}{2}} + (1 + 2 \sigma_1^2 + a \sigma_1^2) \sqrt{\frac{2a}{\pi}} \exp \left(-\frac{a}{2} \right).$$

Since $B = \operatorname{erfc} \sqrt{\frac{a}{2}}$ (again easily found from the marginal distribution) we have

$$\langle E_N^2 E_n^2 \rangle_a = 1 + 2\sigma_1^2 + [1 + \sigma_1^2 (2 + a)] Q \text{ with}$$

$$Q = \sqrt{\frac{2a}{\pi}} \exp \left(-\frac{a}{2} \right) / \operatorname{erfc} \left(\sqrt{\frac{a}{2}} \right).$$

Substitution of the moments listed in Table 1 in (4) gives $R_2(a)$ for correct models [type $(n, 0)$].

In the space group $P1$ we find

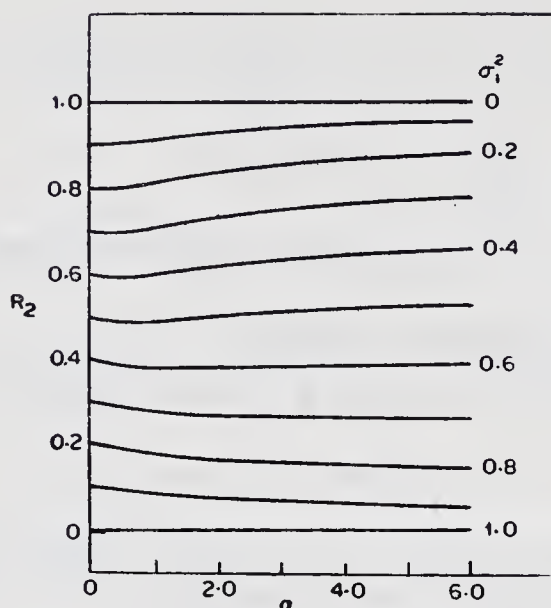
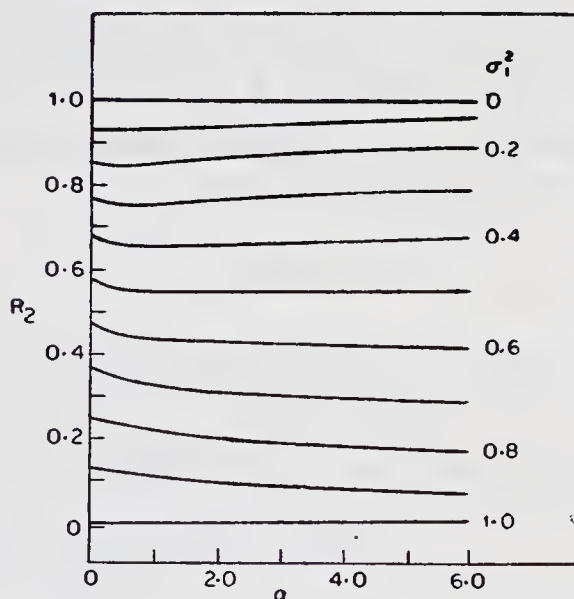
$$R_2(a) = \{a^2 (\sigma_1^8 - 2\sigma_1^4 + 1) + 2a (-\sigma_1^8 + 2\sigma_1^6 - \sigma_1^4 - \sigma_1^2 + 1) + 2(1 - \sigma_1^2)\} \star \{a^2 + 2a + 2\}^{-1}. \quad (11)$$

The behaviour of R_2 for correct models in the space group $P\bar{1}$ is algebraically given by

$$R_2(a) = 1 + \left\{ -\sigma_1^2 (2 + \sigma_1^2) \operatorname{erfc} \sqrt{\frac{a}{2}} + \sigma_1^2 \left[-2 - 4\sigma_1^2 + 6\sigma_1^4 - 3\sigma_1^6 + \sigma_1^2 (\sigma_1^4 - 2) a \right] \sqrt{\frac{2a}{\pi}} \exp \left(-\frac{a}{2} \right) \right\} \star \left[3 \operatorname{erfc} \sqrt{\frac{a}{2}} + (3 + a) \sqrt{\frac{2a}{\pi}} \exp \left(-\frac{a}{2} \right) \right]^{-1}. \quad (12)$$

The functional behaviour of R_2 for correct models in $P1$ and $P\bar{1}$ is depicted in Fig. 3 and Fig. 4 respectively.

From a crystallographic point of view the most interesting series of models are the correct ones $(n, 0)$. The correspondence between theory and experiment is illustrated in Tables 2 and 3. As example of a non-

Fig. 3. R_2 for the related case in space group $P1$.Fig. 4. R_2 for the related case in space group $P\bar{1}$.

centrosymmetric structure we used ammonium-hydrogen l-malate (Versichel, Van de Mierop & Lenstra 1978), space group $P2_12_12_1$, maximum Bragg-angle $\theta = 27^\circ$ (MoK α). As centrosymmetric struc-

Table 2. *Comparison of experimental R_2 values with the theoretical ones as a function of the threshold 'a' for a non-centrosymmetric structure.*

σ_1^2 measures the size of the known, correct structure fragment. Here $\sigma_1^2 = 0.9$ corresponds to a model with 9 out of 10 atoms in the molecule. The asterisk indicates a local minimum in R_2 as a function of a .

a	$\sigma_1^2 = 0.9$		$\sigma_1^2 = 0.5$	
	$100 \times R_2$ (exp)	$100 \times R_2$ (theor)	$100 \times R_2$ (exp)	$100 \times R_2$ (theor)
0.00	9.44	10.00	54.10	50.00
0.20	9.26	9.76	52.91	49.08
0.40	9.11	9.44	52.80	48.65
0.60	8.95	9.08	52.38*	48.53*
0.80	8.94	8.73	52.62	48.58
1.00	8.74	8.40	52.78	48.75
1.20	8.43	8.09	53.48	48.97
1.40	7.49	7.81	54.31	49.22
1.60	7.45	7.56	54.43	49.48
1.80	7.44	7.33	54.65	49.75
2.00	7.30	7.12	54.93	50.00
3.00	6.10	6.33	52.92	51.10
4.00	5.66	5.82	52.00	51.92
5.00	5.56	5.46	52.34	52.53
6.00	5.81	5.20	55.00	53.00

ture we used cis, cis-4, 6-dimethyl trimethylenesulfite (space group $P2_1/c$; Petit, Lenstra and Geise, 1978), in which all actual atoms were replaced by, say nitrogen atoms, to obtain an equal-atom structure.

The correspondence between theory and practice is very satisfactory.

3. The general model (g, f)

Suppose we have calculated a tentative electron-density distribution in which n maxima are found. With respect to the observed N -atom structure this tentative model contains g atoms at their correct position, but f atoms ($n-g=f$) are totally misplaced. The question is clear: 'Are we able to calculate the

Table 3. *Comparison of experimental R_2 values with the theoretical ones as a function of the threshold 'a' for a centrosymmetric structure.*

σ_1^2 measures the size of the known, correct structure fragment, e.g. $\sigma_1^2 = 0.11$ corresponds here to a model with 1 out of the 9 atoms in the molecule. NO represents the number of reflections taken in the calculation of R_2 . $\theta_{\max} = 30^\circ$ (MoK α).

a	NO	$\sigma_1^2 = 0.11$		$\sigma_1^2 = 0.56$	
		$100 \times$ R_2 (exp)	$100 \times$ R_2 (theor)	$100 \times$ R_2 (exp)	$100 \times$ R_2 (theor)
0.0	1561	93.92	92.18	50.86	52.68
0.2	957	93.83*	91.98*	49.45	50.60
0.4	753	93.87	92.07	49.30	49.87
0.6	621	93.98	92.23	49.08	49.40
0.8	532	94.20	92.40	48.83	49.05
1.0	455	94.86	92.59	48.26	48.79
1.2	387	95.00	92.78	48.16	48.59
1.4	344	95.02	92.96	48.15	48.43
1.6	308	95.65	93.13	47.86	48.30
1.8	264	95.71	93.30	47.84	48.19
2.0	240	95.83	93.45	47.78	48.10
3.0	149	96.30	94.12	47.70	47.84
4.0	86	96.61	94.61	47.68	47.72
5.0	58	96.64	95.00	47.68	47.66

R_2 value theoretically for a model (g, f) ?'. The answer to this question is positive. The procedure to predict $R_2 [a, (g, f)]$ is in fact quite simple. All we have to do is to evaluate again the moments indicated in (4). As an example we will discuss one simple term, notably $\langle E_n^2 \rangle_a$.

The normalised intensity for our n -atom model is given by:

$$E_n^2 = \frac{1}{n} \left[\left(\sum_{j=1}^n \cos 2\pi \mathbf{H} \cdot \mathbf{r}_j \right)^2 + \left(\sum_{j=1}^n \sin 2\pi \mathbf{H} \cdot \mathbf{r}_j \right)^2 \right].$$

Handling normalised intensities, any subset of atoms—in our nomenclature N , n , g or f —one always has $\langle E_N^2 \rangle = \langle E_n^2 \rangle = \langle E_g^2 \rangle = \langle E_f^2 \rangle = 1$. So describing the structure as a system of equal point atoms the

scattering power of each atom is inversely proportional to the square root of the number of atoms one uses in the E calculation. Then the normalised intensity of our n -atom model can be written as

$$E_n^2 = \frac{g}{n} E_g^2 + \frac{f}{n} E_f^2.$$

Since E_g and E_f are not interrelated, a simple averaging gives

$$\langle E_n^2 \rangle = \frac{g}{n} \langle E_g^2 \rangle + \frac{f}{n} \langle E_f^2 \rangle = \frac{g}{n} \langle E_g^2 \rangle + \frac{f}{n}$$

Introducing the threshold a on the observed intensities does not influence $\langle E_f^2 \rangle$ at all. So

$$\langle E_n^2 \rangle_a = \frac{g}{n} \langle E_g^2 \rangle_a + \frac{f}{n}. \quad (13)$$

Analogously we obtain:

$$\langle E_N^2 E_n^2 \rangle_a = \frac{g}{n} \langle E_N^2 E_g^2 \rangle_a + \frac{f}{n} \langle E_N^2 \rangle_a \quad (14)$$

$$\langle E_n^4 \rangle_a = \frac{g^2}{n^2} \langle E_g^4 \rangle_a + \frac{f^2}{n^2} \langle E_f^4 \rangle + \frac{agf}{n^2} \langle E_g^2 \rangle_a, \quad (15)$$

where $a=4$ or 6 for the space group $P1$ and $P\bar{1}$ respectively.

Substitution of these moments in (4) gives $R_2[a, (g, f)]$ algebraically. The values of $\langle E_g^4 \rangle_a$, etc are directly available in Table 1 using $\sigma_1^2 = g/N$. To calculate R_2 we substitute these moments in (4), where due to the size n of our model σ_1^2 remains n/N .

This description was verified using computer-simulated experiments. Typical examples for $P1$ and $P\bar{1}$ are listed in Table 4.

The experimental $\langle R_2 \rangle$ - values given are the experimental averages of 200 R_2 -values generated for 200 different structures each containing 100 atoms in the unit cell and described by 2000 reflections. Convergence showed that R_2 averaged over 200 structures is stable up to the third digit.

Table 4. Comparison of 'experimental' R_2 -values against the theoretical ones as a function of the threshold " a " for a non-centrosymmetric and centrosymmetric structure. 'Standard deviations' of the experimental $\langle R_2 \rangle$ -values are shown in parentheses.

Model	Space group $P1$				(60, 0)				(0, 60)			
	a	$\langle R_2 \rangle$	R_2^{theor}	$\langle R_2 \rangle$	R_2^{theor}	$\langle R_2 \rangle$	R_2^{theor}	$\langle R_2 \rangle$	R_2^{theor}	$\langle R_2 \rangle$	R_2^{theor}	R_2^{theor}
Space group $P1$												
	0.0	0.747(35)	0.748	0.667(18)	0.668	0.399(11)	0.400	0.755(17)	0.757			
	1.0	0.485(21)	0.487	0.595(17)	0.597	0.378(11)	0.379	0.661(13)	0.663			
	2.0	0.448(22)	0.449	0.626(21)	0.629	0.379(15)	0.381	0.710(16)	0.711			
	3.0	0.450(30)	0.452	0.658(29)	0.662	0.382(22)	0.385	0.759(22)	0.760			
	4.0	0.456(49)	0.461	0.682(44)	0.687	0.386(35)	0.388	0.795(30)	0.797			
	5.0	0.462(80)	0.471	0.697(65)	0.706	0.390(54)	0.391	0.820(46)	0.825			
Space group $P\bar{1}$												
	0.0	0.995(60)	0.995	0.842(23)	0.837	0.482(17)	0.480	0.955(24)	0.954			
	1.0	0.640(35)	0.642	0.707(18)	0.706	0.440(16)	0.439	0.783(16)	0.784			
	2.0	0.578(32)	0.582	0.710(20)	0.708	0.429(18)	0.429	0.796(16)	0.796			
	3.0	0.556(35)	0.559	0.722(25)	0.720	0.424(21)	0.424	0.819(17)	0.818			
	4.0	0.547(42)	0.549	0.736(31)	0.731	0.419(26)	0.420	0.840(20)	0.839			
	5.0	0.543(54)	0.545	0.748(38)	0.741	0.418(34)	0.418	0.859(24)	0.856			

4. The moments of R_2

To decide whether a tentative model has to be accepted or not we have now a theoretical R_2 value at our disposal. The disadvantage is clear: point estimators are of limited use. A proper statistical decision can only be formulated if the distribution $P(R_2)$ is known. Direct algebraic derivation of this function is not possible. However, an indirect approach appeared possible, viz. the enumeration of all moments of R_2 . For brevity we will only deal with models of the type $(n, 0)$ and $(0, n)$ in one single space group, namely $P1$ and exclude effects due to threshold.

In the previous sections the averages $\langle E^q \rangle_H$ were replaced by intensity distribution averages. In the derivation of the Wilson distribution functions one takes the atomic coordinates as primitive variables. This means that $\langle E^q \rangle_H$ is replaced by $\langle E^q \rangle_r$. For infinite data sets these averages are indeed equal. Unfortunately this equality becomes an approximation handling finite data sets. This is one reason why the logic used in the previous sections cannot be used beyond the level of a first central moment. A second problem we have to cope with is that R_2 , etc. are related to one single observed structure of size N instead of to some sort of average N -atom structure. This means that $P(E_N)$ is no longer representative for our actual problem; this distribution has to be replaced by $\{E_N(H)\}_K$ as a set of fixed quantities describing our observed structure. For reasons of mathematical simplicity we take the finite data set as an aselect subset of the whole, infinite data set. As a consequence equation (4) does not hold anymore. It has to be replaced in terms of the above outlines by:

$$\begin{aligned} \langle R_2; \{E_N(H)\}_K \rangle = 1 + \sigma_1^4 \frac{\sum_H \langle E_n^4; \{E_N(H)\}_K \rangle}{\sum_H E_N^4(H)} \\ - 2\sigma_1^2 \frac{\sum_H E_N^2 \langle E_n^2; \{E_N(H)\}_K \rangle}{\sum_H E_N^4(H)}, \quad (16) \end{aligned}$$

with $\{E_N\}_H \subset \{E_N\}_K$. To calculate the moments indicated in (16) we have to know the conditional probability function $P[E_n(H); \{E_N(H)\}_K]$. The formulation of this conditional probability has to reflect the fact that the proposed n -atom model is correct or incorrect.

4.1. *Incorrect structures of type (0, n)*

Since the correlation between every proposed n -atom model and the observed structure is absent, we have

$$\langle E_n^q(H); \{E_N(H)\}_K \rangle_{r_n} = \langle E_n^q(H) \rangle_{r_n} \quad (17)$$

Using the asymptotic Wilson distribution $2E_n \exp[-E_n^2]$ the right-hand side of (17) can be enumerated as

$$\langle E_n^q(H) \rangle_{r_n} = \Gamma\left(\frac{q}{2} + 1\right). \quad (18)$$

Substitution of these moments in (16) gives

$$\langle R_2 \rangle_{r_n} = 1 + \sigma_1^4 \frac{\sum_H 2}{\sum_H E_N(H)^4} - 2 \sigma_1^2 \frac{\sum_H E_N(H)^2}{\sum_H E_N(H)^4}. \quad (19)$$

This function describes R_2 for a particular set of structure factors of the observed N-atom structure. It is also evident that the introduction of a threshold a is not very difficult. Equation (6) has to be a limiting case of (19). This can be easily shown. When we are interested in an overall picture of R_2 , applicable to any observed N-atom structure, we need to average over all observable structures of N-atoms. We then find

$$\langle \langle R_2 \rangle_{r_n} \rangle_{r_N} = 1 + \sigma_1^4 - \sigma_1^2,$$

which is indeed equal to (6) with a threshold $a=0$.

This simple experiment suffices to show that the results of the previous sections are a special case in the improved concept developed in this part.

The central moments μ_q of R_2 are given by

$$\mu_q = \langle (R_2 - \langle R_2 \rangle)^q \rangle.$$

In terms of the geometrical moments μ'_q the central moments are algebraically expressed by

$$\mu_q = \sum_{j=0}^q \binom{q}{j} (-1)^{q-j} \mu'_j \mu'_1{}^{(q-j)}.$$

After some manipulation we find

$$\begin{aligned} \mu_q = & \sum_{i=0}^q \sum_{j=0}^{q-i} \sum_{k=0}^{q-i-j} \frac{q!(-1)^{k+i} 2^{j+k+i}}{i!j!k!(q-i-j-k)!} \sigma_1^{2(2q-j-k)} \star \\ & \sum_H E_N^{2(j+k)} \langle E_n^{2(2q-2i-2j-k)} \rangle / \left(\sum_H E_N^4 \right)^q. \end{aligned}$$

From (18) we know

$$\langle E_n^{2(2q-2i-2j-k)} \rangle = (2q-2i-2j-k)!$$

Let $q=2$, i.e. we calculate $\sigma^2(R_2)$. The resulting expression reads

$$\begin{aligned} \langle \sigma^2(R_2) \rangle_{\mathbf{r}_n} = & \left(4\sigma_1^4 \sum_H E_N^4 - 16\sigma_1^6 \sum_H E_N^2 \right. \\ & \left. + 20\sigma_1^8 \sum_H 1 \right) / (\Sigma E_N^4)^2, \end{aligned}$$

where $1 = E_N^\circ$. We see that $\sigma^2(R_2)$ is inversely proportional to the number of observations, which corroborates with one's intuitive expectation. Of more direct use is the fact that for practical purposes $P(R_2)$ can be regarded as a Gaussian distribution. In table 5 the values of R_2 , $\sigma(R_2)$ and the skewness ($\gamma_1 = \mu_3/\sigma_3$) are listed for a structure of 500 atoms with 10 000 reflections. The numbers are calculated using the additional averaging over \mathbf{r}_N in order to obtain an overall picture rather than a specific one. The

Table 5. *The theoretical values of R_2 , $\sigma(R_2)$ and μ_3/σ^3 for incorrect models of different size n . $P(R_2)$ is nearly Gaussian, because the skewness $\gamma_1 \sim 0$.*

n	Threshold $a = 0$			Threshold $a = 1$		
	$\langle R_2 \rangle$	σR_2	γ_1	$\langle R_2 \rangle$	σR_2	γ_1
0	1.0000	0.0000	—	1.0000	0.0000	—
25	0.9525	0.0007	-0.0424	0.9610	0.0007	-0.0463
50	0.9100	0.0013	-0.0421	0.9240	0.0014	-0.0454
75	0.8725	0.0018	-0.0414	0.8890	0.0020	-0.0446
100	0.8400	0.0024	-0.0399	0.8560	0.0025	-0.0439
125	0.8125	0.0029	-0.0372	0.8250	0.0030	-0.0533
150	0.7900	0.0034	-0.0329	0.7960	0.0035	-0.0427
175	0.7725	0.0039	-0.0268	0.7690	0.0039	-0.0420
200	0.7600	0.0044	-0.0187	0.7440	0.0043	-0.0412
225	0.7525	0.0050	-0.0090	0.7210	0.0046	-0.0398
250	0.7500	0.0056	0.0018	0.7000	0.0049	-0.0379
275	0.7525	0.0063	0.0130	0.6810	0.0053	-0.0348
300	0.7600	0.0071	0.0241	0.6640	0.0056	-0.0304
325	0.7725	0.0080	0.0343	0.6490	0.0059	-0.0242
350	0.7900	0.0090	0.0433	0.6360	0.0063	-0.0158
375	0.8125	0.0101	0.0511	0.6250	0.0067	-0.0059
400	0.8400	0.0113	0.0576	0.6160	0.0071	0.0074
425	0.8725	0.0126	0.0629	0.6090	0.0075	0.0219
450	0.9100	0.0141	0.0671	0.6040	0.0081	0.0376
475	0.9525	0.0156	0.0705	0.6010	0.0087	0.0539
500	1.0000	0.0173	0.0731	0.6000	0.0093	0.0699

kurtosis $[(\mu_4/\sigma^4)-3]$ is ca. -2.99 , which means that $P(R_2)$ is platykurtic.

4.2. Correct models of type $(n, 0)$

To calculate R_2 using (16) we need to know the intensity moments $\langle E_n^2(H); \{E_N(H)\}_K \rangle$. These moments are, however, too complicated to obtain without an additional assumption. Let us ignore the correlation, if any, between the intensities of the reflections H and K . Then the required intensity moments can be simplified to: $\langle E_n^q(H); E_N(H) \rangle$. As far as practical results are decisive in underlining a theoretical concept, the final tests show that this approximation holds in its consequences.

To calculate $\langle E_n^q; E_N \rangle$ we need to construct the conditional probability function $P(E_n; E_N)$. Fortunately for correct models the conditional probability function $P(E_N; E_n)$ has been derived (Srinivasan & Parthasarathy, 1976). This function is given by

$$P(E_N; E_n) = \frac{2E_N\sigma_N^2}{\sigma_N^2 - \sigma_n^2} \exp \left[-\frac{\sigma_N^2 E_N^2 + \sigma_n^2 E_n^2}{\sigma_N^2 - \sigma_n^2} \right]$$

$$I_0 \left(\frac{2\sigma_N\sigma_n E_N E_n}{\sigma_N^2 - \sigma_n^2} \right),$$

$$\text{with } \sigma_N^2 = \sum_{j=1}^N f_j^2 \text{ and } \sigma_n^2 = \sum_{j=1}^n f_j^2.$$

To get the necessary distribution $P(E_n; E_N)$ we apply the theorem of Bayes

$$P(E_n; E_N) = \frac{P(E_N; E_n)P(E_n)}{P(E_N)}.$$

Substitution of the relevant distributions at the right-hand side gives

$$P(E_n; E_N) = \frac{2E_n \sigma_N^2}{\sigma_N^2 - \sigma_n^2} \exp \left[- \frac{\sigma_N^2 E_n^2 + \sigma_n^2 E_N^2}{\sigma_N^2 - \sigma_n^2} \right] \\ I_0 \left(\frac{2\sigma_N \sigma_n E_N E_n}{\sigma_N^2 - \sigma_n^2} \right).$$

The moments of this distribution are given by

$$\langle E_n^q; E_N \rangle = \Gamma \left(\frac{q}{2} + 1 \right) \sigma_n^q {}_1F_1 \left(-\frac{q}{2}; 1; -\frac{\sigma_n^2}{\sigma_N^2} E_N^2 \right), \quad (20)$$

$$\text{where } {}_1F_1(x; y; z) = 1 + \frac{x}{y} z + \frac{x(x+1)}{y(y+1)} \frac{z^2}{2!} \dots$$

a so-called hypergeometric function.

Substitution of the relevant moments in (16) gives

$$\langle R_2 \rangle_{r_n} = \{ (1 - \sigma_1^4) \Sigma E_N^4 + 2 \sigma_1^2 (1 - 2 \sigma_1^4) (\sigma_1^2 - 1)^* \\ \Sigma E_N^2 + 2 \sigma_1^4 (1 - \sigma_1^2) \Sigma 1 \}^* \{ \Sigma E_N^4 \}^{-1}.$$

For correct models we calculated the q^{th} central moment too. After some manipulations we find

$$\mu_q = \left\{ \sum_{i=0}^q \sum_{j=0}^{q-i} \sum_{k=0}^{q-i-j} \sum_{l=0}^{q-i-j-k} \right. \\ = \frac{q! (-1)^{l+i} 2^{l+i+j}}{i! j! k! l! (q-i-j-k-l)!} \sigma_1^{2(2q-j-l)} * \\ (1 - \sigma_1^2)^{j+2i} (2 - \sigma_1^4)^k (1 - 2\sigma_1^4)^j \\ \left. \Sigma E_N^{2(2k+j+l)} \langle E_n^{2(2q-2i-2j-2k-l)}; E_N \rangle \right\}^* \\ \left\{ \left(\sum_H E_N^4 \right)^q \right\}^{-1}.$$

Since only the even moments of E_n are required, the confluent hypergeometric function [see (20)] can be replaced by a finite polynomial. Therefore we have

$$\langle E_n^{2m}; E_N \rangle = \sum_{p=0}^m \frac{m! m!}{p! p! (m-p)!} \sigma_1^{2p} (1 - \sigma_1^2)^{m-p} E_N^{2p}$$

Taking $q=2$ we find $\sigma^2(R_2)$

$$\begin{aligned} \sigma^2(R_2) = & \{ (1 - \sigma_1^2) (8\sigma_1^{14} - 16\sigma_1^{10} + 8\sigma_1^6) \Sigma E_N^6 \\ & + (1 - \sigma_1^2)^2 (52\sigma_1^{12} - 48\sigma_1^8 + 4\sigma_1^4) \Sigma E_N^4 \\ & + (1 - \sigma_1^2)^3 (80\sigma_1^{10} - 16\sigma_1^{16}) \Sigma E_N^2 \\ & + 20\sigma_1^8 (1 - \sigma_1^2)^4 \Sigma 1 \} / (\Sigma E_N^4)^2. \end{aligned}$$

In Table 6 R_2 , $\sigma(R_2)$ and the skewness are given for an 'overall' structure of 500 atoms characterised with 10 000 reflections. The value of the kurtosis $(\mu_4/\sigma^4 - 3)$ is again ca -2.99 . Once more we can decide that $P(R_2)$ for correct structure models can be regarded for practical purposes as a Gaussian distribution.

Using the concept of a Gaussian function the first two moments $\langle R_2 \rangle$ and $\sigma(R_2)$ are sufficient to describe $P(R_2)$. We tested this in a practical situation, taking N-acetyl-allohydroxy-L-proline lactone (Lensa, Petit & Geise 1979) as an example. The asymmetric part of the unit cell was taken as a P1 structure in which the original $C_7H_9NO_3$ -moiety was replaced by X_{11} , where X has a scattering power of $1/\sqrt{11}$. Each model of n -atoms was calculated in all possible $\binom{n}{11}$ ways. This gave for every model of n -atoms three final R -values, viz. its minimum, its maximum and its averaged values. The agreement between theory and practice is illustrated in Figs. 5, 6 & 7.

Inspection of these functions reveals an astonishing property of $\sigma(R_2)$. Its numerical value for models of the type $(n, 0)$ hardly changes when $\sim 80\%$ of the data,

Table 6. *The expectation values for R_2 , $\sigma(R_2)$ and $\gamma_1 = \mu_3/\sigma^3$ as a function of the size n of the correct trial structure*

n	Threshold $a = 0$			Threshold $a = 1$		
	$\langle R_2 \rangle$	σ_{R_2}	γ_1	$\langle R_2 \rangle$	σ_{R_2}	γ_1
0	1.0000	0.0000	—	1.0000	0.0000	—
25	0.9500	0.0007	-0.0457	0.9581	0.0008	-0.0491
50	0.9000	0.0015	-0.0450	0.9128	0.0016	-0.0474
75	0.8500	0.0023	-0.0423	0.8646	0.0024	-0.0441
100	0.8000	0.0028	-0.0383	0.8141	0.0032	-0.0398
125	0.7500	0.0036	-0.0337	0.7617	0.0039	-0.0350
150	0.7000	0.0042	-0.0284	0.7080	0.0045	-0.0298
175	0.6500	0.0047	-0.0228	0.6533	0.0051	-0.0242
200	0.6000	0.0051	-0.0169	0.5981	0.0055	-0.0183
225	0.5500	0.0054	-0.0107	0.5427	0.0058	-0.0120
250	0.5000	0.0056	-0.0042	0.4875	0.0059	-0.0053
275	0.4500	0.0056	0.0026	0.4328	0.0059	0.0017
300	0.4000	0.0055	0.0095	0.3789	0.0058	0.0090
325	0.3500	0.0052	0.0167	0.3260	0.0055	0.0167
350	0.3000	0.0048	0.0240	0.2744	0.0051	0.0247
375	0.2500	0.0043	0.0313	0.2242	0.0045	0.0329
400	0.2000	0.0036	0.0387	0.1757	0.0038	0.0412
425	0.1500	0.0028	0.0458	0.1289	0.0029	0.0494
450	0.1000	0.0019	0.0523	0.0840	0.0020	0.0571
475	0.0500	0.0010	0.0575	0.0410	0.0010	0.0634
500	0.0000	0.0000	—	0.0000	0.0000	—

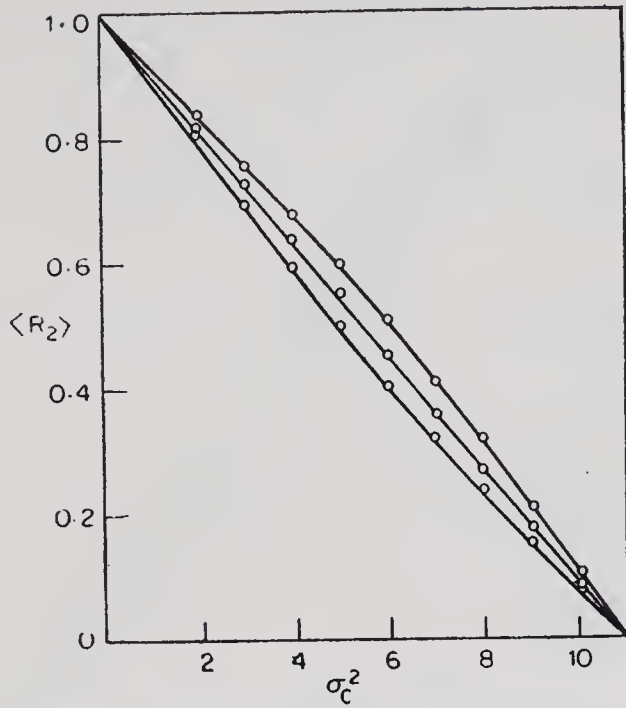


Fig. 5. $0 \leq E^2$ Experimental points are given by circles, solid lines represent theoretical values.

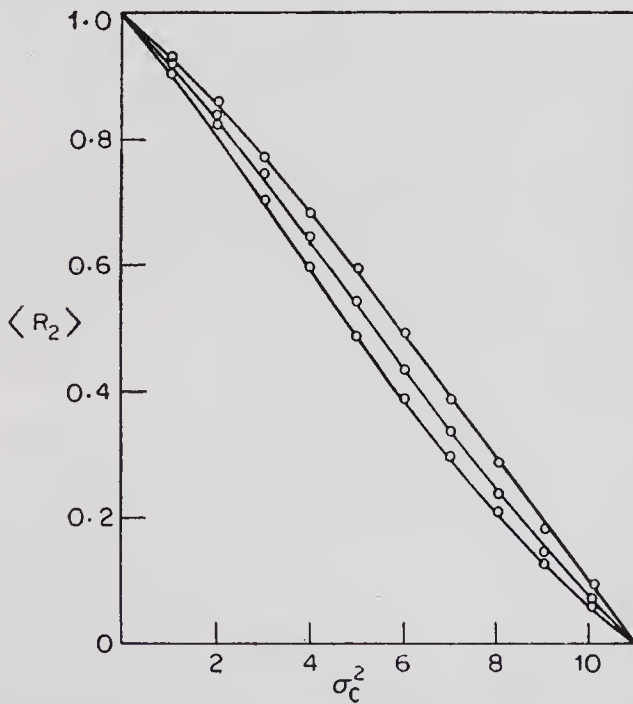


Fig. 6. Data set $1 \leq E^2$ Experimental points are given by circles, solid lines represent theoretical values.

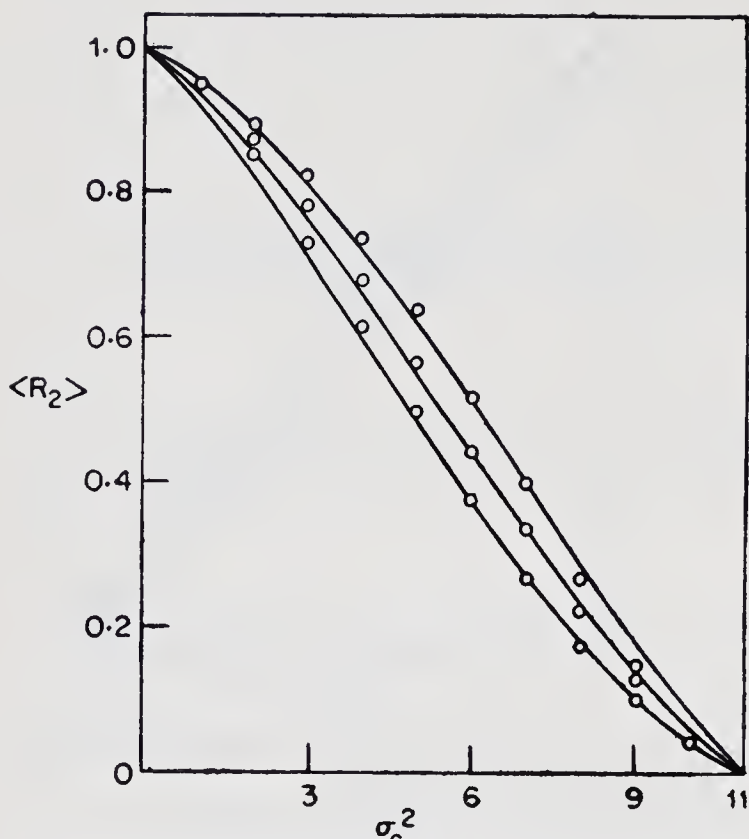


Fig. 7. Data set $2 \leq E^2$ Experimental points are given by circles, solid lines represent theoretical values.

notably the low intensities, are omitted from the calculations. This surprising feature makes it look profitable to use only the large intensities to test the reliability of a tentative structure. To prevent over-optimistic applications, one has to look at the consequences of this behaviour. This is, at least at one dominant point, worked out in more detail by Petit & Lenstra (1982).

R_2 and its behaviour for correct models is sometimes not at all simple. One expects that R_2 will have values between 1 and 0. This is not true! Let us take those reflections, which in direct methods are used to calculate Ψ -zero. Using the same structure that served to calculate the experimental points in the previous Figs. 5, 6 & 7 we obtain Fig. 8 using error-free E_N^2 -values. Again theory and experiment

show the same features, though now qualitative and not quantitative. At present this is a matter of further investigations. The main lesson at this point is that it summons the question 'Is psi-zero truly a nice indicator function?'. Looking at Fig. 8 our present guess is: 'It is not!'. If it is, please explain why.

5. A first application of the theory: rotation search

The purpose of a rotation search is to locate a known molecular fragment in an orientation [C] in a unit cell coinciding with the one of the observed N-atom structure. The rotation search is always performed in the triclinic space groups $P1$ and $P\bar{1}$. This means that rotation search is a fine example of the situations $(0, n)$ and $(n, 0)$.

In reciprocal space the measure of fit between model and observed structure is given by:

$$R(C) = \sigma_1^2 \sum_H E_N^2 E_n^2(C) \quad (21)$$

which is simply the double product term existing in the definition of R_2 .

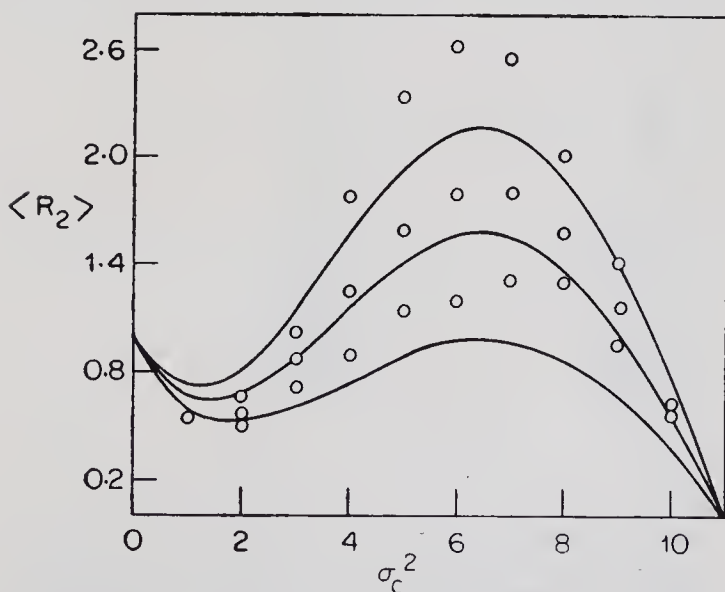


Fig. 8. Data set $0 < E_N^2 < 0.5$ Experimental points are given by circles, solid lines represent theoretical values,

In Fig. 9 two histograms are shown, in which ammonium hydrogen malate (Versichel, van de Mierop & Lenstra, 1978) served as test compound. Both histograms are practically the same in spite of the difference in the sampling technique. R_{\max} corresponds to the correct orientation. All other R -values correspond to incorrect orientations. Looking at these histograms we see that $P(R)$ is not Gaussian. There is definitely some tailing in the direction of R_{\max} . This tailing effect can be easily explained in terms of partially correct oriented models. By this we mean that the chemically known fragment is well oriented with exception for some of its substituents. Using the terminology of §3, we have

$$R(g,f)_a = \text{NO} \left[\frac{g}{n} \{ \sigma_1^2 a^2 + (1 + \sigma_1^2) a + 1 + \sigma_1^2 \} + \frac{f}{n} (1 + a) \right] e^{-a \sigma_1^2} \quad (22)$$

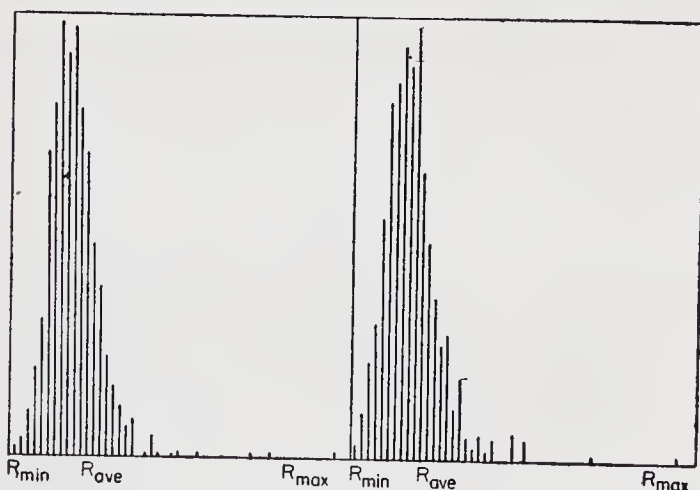


Fig. 9. Histograms of experimental R -values in a rotation search of ammonium hydrogen (1)-malate with a 10-atom search fragment using the Eulerian angle sampling technique (left) and the optimal sampling technique of Lattman (1972). To make comparison easier, the number of sampling points in both techniques has been scaled to the same value.

for the space group $P1$, and

$$R(g, f)_a = \text{NO} \left[\frac{g}{n} \{1 + 2 \sigma_1^2 + (1 + (2 + a) \sigma_1^2) Q\} + \frac{f}{n} (1 + Q) \right] \text{erfc} \sqrt{\frac{a}{2}} \sigma_1^2, \quad (23)$$

for the space group $P\bar{1}$, where $\sigma_1^2 = g/N$ and Q as in Table 1 and NO is the effective number of observations.

The proper formulae for $R(n, 0)$ and $R(0, n)$ are easy to obtain. For the space group $P1$ we have:

$$R(0, n)_a = \text{NO} (1 + a) e^{-a} \sigma_1^2,$$

$$R(n, 0)_a = \text{NO} [1 + \sigma_1^2 + (1 + \sigma_1^2) a + \sigma_1^2 a^2] e^{-a} \sigma_1^2.$$

A numerical verification is summarised in Table 8, where $R(n, 0)/R(0, n)$ is given as well as the experimental ratio of $R_{\text{max}}/R_{\text{ave}}$ in which R_{ave} is the experimental average of R over all data points. Let us concentrate on incorrectly oriented models. Following the description in §4

$$\sigma^2(R) = \sigma_1^4 \sum_H E_N^4 \text{ for } R = \sigma_1^2 \sum E_N^2 E_n^2. \quad (24)$$

Table 7. *Experimental approach to R_f -distribution for ammonium hydrogen (1)-malate with a 10-atom search fragment. $Z = R_{\text{ave}} - R_{\text{min}}$.*

a	Z	Z/σ σ of (24)	Z/σ σ of (25)
0.0	70.6	2.60	3.29
1.0	74.7	2.76	3.64
2.0	61.3	2.48	3.48
3.0	66.4	2.87	4.76
4.0	55.0	2.61	5.26
5.0	46.2	2.45	6.10
6.0	38.1	2.37	7.15

Consequently $\sigma^2(R)$ for a particular structure to be determined, can be calculated prior to an actual rotation search. When $P(R)$ can be taken as Gaussian, this means that $Z=(R_{\text{ave}}-R_{\text{min}})$ will be roughly 3 times $\sigma(R)$ for the ca. 1400 sampling points depicted in Fig. 9. In Table 7 the numerical value of Z is given as a function of the threshold a . Using (24) the ratio Z/σ is also shown with the threshold as parameter. With our present knowledge we are able to formulate the rotation search in such terms, that the correct orientation can be found at minimal computer costs due to an 'error-analysis' prior to the actual calculations.

The charm of this example is also that it allows us to illustrate explicitly the consequences of gaining an overall picture. To get an overall picture of $\sigma(R)$ we replace ΣE_N^4 in (24) by its statistical value. We then replace

$$\langle \sigma^2(R) \rangle_{\mathbf{r}_n} \text{ by } \langle \langle \sigma^2(R) \rangle_{\mathbf{r}_n} \rangle_{\mathbf{r}_N}.$$

Table 8. *Comparison of theory and experiment in the non-centrosymmetric case as a function of the threshold, with a search fragment of size $\sigma_1^2=0.234$.*

a	NO		$R(n, 0)/R(0, n)$	
	actual	theory	exp.	theory
0.0	3677	3677	1.28	1.23
0.2	2770	3010	1.29	1.24
0.4	2153	2465	1.31	1.26
0.6	1757	2018	1.33	1.29
0.8	1454	1652	1.35	1.32
1.0	1188	1353	1.38	1.35
1.2	976	1107	1.42	1.39
1.4	795	907	1.46	1.43
1.6	672	742	1.48	1.46
1.8	560	608	1.50	1.50
2.0	454	498	1.53	1.55
3.0	278	183	1.61	1.76
4.0	159	67	1.71	1.98
5.0	97	25	1.82	2.21

(24) then becomes

$$\sigma^2(R) = \sigma_1^4 \text{NO } e^{-a} (a^2 + 2a + 2) \text{ for } P1 \quad (25)$$

and

$$\sigma^2(R) = \sigma_1^4 \text{NO } \operatorname{erfc} \sqrt{\frac{a}{2}} [3 + (3 + a) Q] \text{ for } P\bar{1}.$$

Taking σ from (25) we also calculated Z/σ . For qualitative purposes the values are useful, because they are of the same magnitude as the ones tailored to our example. For applications, it is clear that (24) is of much more value than (25). The large discrepancies in ammonium hydrogen 1-malate between σ (24) and σ (25) are due to the fact that $P(E_N)$ has a centric character. This is shown in Table 8, where $\text{NO } e^{-a}$ is compared with the actual number of E^2 -values above a threshold a .

The author wishes to thank his co-workers W. van de Mieroop, G. H. Petit, W. van Havere, M. van Poucke, J. van Loock for their essential contributions to this paper. Prof. Dr. H. J. Geise is gratefully acknowledged for his critical reading of the try-outs.

References

- ABRAMOWITZ, M. & STEGUN, I. A. (1968). *Handbook of Mathematical Functions*. New York: Dover Publications.
- VAN HAVERE, W. & LENSTRA, A. T. H. (1980). *Abstracts Sixth European Cryst. Meeting, Barcelona*, p. 156.
- LATTMAN, E. E. (1972). *Acta Cryst.* B28, 1065–1068.
- LENSTRA, A. T. H. (1973). Ph.D. Thesis, State University of Utrecht, The Netherlands.
- LENSTRA, A. T. H. (1974). *Acta Cryst.* A30, 363–369.
- LENSTRA, A. T. H. (1979). *Bull. Soc. Chim. Belg.*, 88, 359–368.
- LENSTRA, A. T. H., PETIT, G. H. & GEISE, H. J. (1979). *Cryst. Struct. Comm.* 8, 1023–1029.
- VAN DE MIEROOP, W. (1979). Ph.D. Thesis (in Dutch), University of Antwerp (U.I.A.), Belgium.

- VAN DE MIEROOP, W. & LENSTRA, A. T. H. (1978). *Acta Cryst.* A34, 860-863.
- PARTHASARATHY, S. & PARTHASARATHI, V. (1972). *Acta Cryst.* A28, 426-432.
- PETIT, G. H. & LENSTRA, A. T. H. (1979). *Abstracts Fifth European Cryst. Meeting, Copenhagen*, 344-345.
- PETIT, G. H. & LENSTRA, A. T. H. (1982), *Acta Cryst.* A38 67-70.
- PETIT, G. H., LENSTRA, A. T. H., & GEISE, H. J. (1978). *Bull. Soc. Chim. Belg.*, 87, 659-666.
- PETIT, G. H., LENSTRA, A. T. H. & VAN LOOCK, J. F. (1981). *Acta Cryst.* A37, 353-360.
- SRINIVASAN, R. & PARTHASARATHY, S. (1976). *Some statistical applications in X-ray crystallography*. Oxford: Pergamon Press.
- TOLLIN, P. & ROSSMAN, M. G. (1966). *Acta Cryst.* 21, 872-876.
- VERSICHEL, W., VAN DE MIEROOP, W. & LENSTRA, A. T. H. (1978). *Acta Cryst.* B34, 2643-2645.
- WILSON, A. J. C. (1969). *Acta Cryst.* B25, 1288-1293.
- WILSON, A. J. C. (1974). *Acta Cryst.* A30, 836-838.
- WILSON, A. J. C. (1976). *Acta Cryst.* A32, 53-56.

Alternatives to Least Squares

BY A. J. C. WILSON

*Department of Physics, University of Birmingham,
Birmingham B15 2TT, England*

Editorial Note

Least-squares adjustment is undoubtedly the commonest method of estimating parameters, but it is not the only one. In crystallography Fourier methods have been much used, and in statistics in general the method of maximum likelihood is probably the main rival of least squares. It was hoped to have a review paper on 'alternatives to least squares' for this symposium. Unfortunately none of those invited was able to accept, though there were contributed papers related to the topic; two of them are reproduced below (pp. 229, 269). It may be useful to try to put some of the problems in perspective.

Naive least-squares gives unit weight to observations. Least-squares adjustment of the observed and calculated intensities is then exactly equivalent to obtaining the best fit between the Patterson density, as represented by the observed intensities, and the Patterson density calculated from the model structure. Statistical fluctuations in the observations do not bias the structural parameters. With certain reservations, it may be said that least-squares refinement based on structure factors is equivalent to making a least-squares fit between the observed and the calculated electron densities, but the parameters obtained would be subject to some bias. Orthodox least-squares uses weights inversely proportional to the estimated standard deviations, and if, as is usual, the estimates depend on actual observations, the resulting parameters are subject to bias, probably not quite negligible for scale and thermal parameters in work of ordinary accuracy, and not quite negligible for

C.S.—15

structural parameters in work of the highest accuracy. Such refinements are equivalent to obtaining a fit between the 'observed' Patterson or electron densities distorted by the statistical weights, and similarly distorted calculated densities. These matters have been discussed in greater detail elsewhere (Wilson 1976*a, b*, 1978). Various non-orthodox weighting schemes have been proposed or used; besides the 'robust/resistant' and others described below, reference may be made to papers by Nielsen (1977), Davis, Maslen & Varghese (1978) and Rees (1978). Bias has not been investigated.

Maximum-likelihood methods depend on a knowledge of, or an assumption about, the distribution function of the statistical fluctuations and other random errors in the observations, whereas least-squares methods depend on little more than a finite variance for these errors. If the distribution function were known, maximum likelihood would give the better estimate, as it incorporates more information. In general the distribution function is not known; all that can be said with reasonable certainty is that for intensity measurements the distribution is only approximately Gaussian, and that large fluctuations are more frequent than would be expected for a Gaussian distribution (for the theory see Wilson, 1980; for empirical evidence see de Boer, this volume, p. 179-186). At least two crystallographic applications of maximum likelihood have been made: Beu (Beu, Musil & Whitney, 1962, and several later papers with various collaborators) used it for lattice-parameter determination, and Price (1979) has proposed to use it for structural parameters. For a Gaussian distribution, as assumed by Beu, Musil & Whitney, maximum likelihood is practically equivalent to least squares (Hamilton, 1964, Bard, 1974). The case for likelihood methods has been persuasively argued by Edwards (1972), and the crystallographic applications have been discussed by Mandel (1980) and Wilson (1980).

References

- BARD, Y. (1974). *Nonlinear Parameter Estimation*. New York: Academic Press, p. 65.
- BEU, K. E., MUSIL, F. J. & WHITNEY, D. R. (1962). *Acta Cryst.* **15**, 1292–1301.
- DAVIS, C. L., MASLEN, E. N. & VARGHESE, J. N. (1978). *Acta Cryst.* **A34**, 371–377.
- EDWARDS, A.W.F. (1972). *Likelihood*. Cambridge: Univ. Press.
- HAMILTON, W. C. (1964). *Statistics in Physical Science*. New York: Ronald Press. 37–42.
- MANDEL, J. (1980). *Accuracy in Powder Diffraction*, pp. 353–359. NBS Special Publication 567. Washington: US Govt Printing Office.
- NIELSEN, K. (1977). *Acta Cryst.* **A33**, 1009–1010.
- PRICE, P. F. (1979). *Acta Cryst.* **A35**, 57–60.
- REES, B. (1978). *Acta Cryst.* **A34**, 254–256.
- WILSON, A. J. C. (1976a). *Acta Cryst.* **A32**, 781–783.
- WILSON, A. J. C. (1976b). *Acta Cryst.* **A32**, 994–996.
- WILSON, A. J. C. (1978). *Acta Cryst.* **A34**, 474–475.
- WILSON, A. J. C. (1980). *Acta Cryst.* **A36**, 929–936.

A Robust/Resistant Technique for Crystal-Structure Refinement

BY W. L. NICHOLSON

*Battelle Pacific Northwest Laboratories,
Richland, WA 99352, U.S.A.*

E. PRINCE

*National Measurement Laboratory, National Bureau
of Standards, Washington, D.C. 20234, U.S.A.*

J. BUCHANAN AND P. TUCKER

*Battelle Pacific Northwest Laboratories,
Richland, WA 99352, U.S.A.*

Abstract

A refinement technique is 'robust' if it works well over a broad class of error distributions in the data, and 'resistant' if it is not strongly influenced by any small subset of the data. Least squares possesses neither property. A more robust/resistant procedure is to minimize, instead of a simple sum of squared differences, a sum of terms of the form $(x^2/2)[1 - (x/a)^2 + (1/3)(x/a)^4]$ for $|x| \leq a$ and $a^2/6$ for $|x| > a$. Here $x = w^{1/2} (|F_0| - |F_c|)/s$, s is a measure of the width of the error distribution based on the results of the previous cycle, and a is a constant chosen so that extreme data do not influence the solution. This function behaves like the sum of squares for small $|x|$, but is constant for large $|x|$, so that the effect of large differences is deemphasized. A least-squares program can easily be modified to perform this more robust/resistant procedure. The modified procedure has been used in a reanalysis of the $D(+)$ - tartaric acid data collected by the Single Crystal Intensity

Project of the International Union of Crystallography [Abrahams, Hamilton & Mathieson (1970), *Acta Cryst.* A26, 1–17). The results show that the technique provides an efficient means for automatic screening of a data set for discrepant data points. It gives results in agreement with the least-squares results for good data sets. If the results do not agree with least squares it suggests systematic effects. A detailed analysis of residuals may identify the problem and help to determine whether the robust/resistant refinement is an improvement.

Introduction

A technique for fitting a theoretical model to a set of experimental data points and estimating the best values of adjustable parameters in the model is said to be 'robust', or, more precisely, 'robust of efficiency', if the parameter estimates have near minimum variance for a broad class of distributions for the errors in experimental data. A technique is 'resistant' if the estimates are not highly dependent on any small subset of the experimental data. To date, all techniques that are robust of efficiency are also resistant. Some data analysts feel that this is inevitable, hence the splice word robust/resistant.

The technique of least squares, the one most commonly used for refining crystal structures, is neither robust nor resistant. Least squares was designed specifically by Gauss for an idealized distribution of errors—an error distribution now referred to as Gaussian. Typically, the errors in experimental crystallographic structure factors are not Gaussian. Inadequacy of the structure factor model induces correlation and bias in residuals, which may propagate as bias in parameter estimates. True precision of experimental data is not known. The net result is that error distributions are usually much longer tailed than Gaussian, and error subsets may be highly correlated.

In recent years, there have been a number of studies of variations of the traditional methods of fitting models to data, to develop techniques that are more robust/resistant than least squares. Tukey (1974) has described the properties such a method should have. The weakness of least squares lies in its great sensitivity to the occurrence of data points widely deviating from their population means with a frequency greatly exceeding that predicted by a Gaussian distribution. A robust/resistant technique treats the body of data in a manner similar to least squares. Wild data are ignored, with a smooth transition of treatment for intermediate situations between these extremes.

In this article, we present a study of the application of robust/resistant techniques to crystal structure refinement, using the data taken on crystals of (*D*+)—tartaric acid for the Single Crystal Intensity Measurement Project of the International Union of Crystallography (Abrahams, Hamilton, & Mathieson, 1970). Our purpose for this study is two-fold: first, to use crystal-structure refinement as a nontrivial test of the robust/resistant approach; and second, to improve crystallographic refinement techniques.

Robust/resistant refinement

Let $y_i = |F_{\text{obs}}(\mathbf{h}_i)|$ ($i = 1, 2, \dots, N$) be a set of N experimentally determined structure amplitudes, where \mathbf{h}_i is a reciprocal lattice vector. We consider the problem of fitting the usual model

$$m_i(\theta) = |F_{\text{calc}}(\mathbf{h}_i, \theta)|$$

where

$$F_{\text{calc}}(\mathbf{h}_i, \theta) = Q E_i \sum_j f_j \exp(2\pi i \mathbf{h}_i \cdot \mathbf{r}_j + \mathbf{h}_i^T \beta_j \mathbf{h}_i). \quad (1)$$

The sum is taken over the set of atoms in the asymmetric unit, Q is a scale factor, E_i an extinction factor and $I = \sqrt{-1}$. For the j -th atom, f_j is the atomic scattering factor, \mathbf{r}_j is the position vector, and β_j is the thermal vibration tensor. θ is a p -dimensional vector of unknown parameters, including scale factor, extinction parameter, atom position coordinates, and atom thermal vibration parameters. More complex models with multiple scaling and extinction parameters, etc., are easy extensions and can be handled in a fashion similar to the discussion here.

A conventional 'full matrix' weighted least-squares refinement fits structure factors to model by selecting θ to minimize

$$\sum_{i=1}^N r_i^2(\theta), \quad (2)$$

where $\mathbf{r}_i(\theta) = \sqrt{w_i} [y_i - m_i(\theta)]$ is the i^{th} standardized residual. Here, w_i is a weighting factor which reflects the precision of the measured structure factor amplitude y_i . Statistical theory suggests that $w_i = 1/\sigma_i^2$, where σ_i^2 is the variance of y_i . In practice, σ_i^2 is unknown and must be estimated. A common estimate among crystallographers is

$$w_i = 1/[\sigma_{si}^2 + (by_i)^2], \quad (3)$$

where σ_{si}^2 is an estimate of the variance due to counting statistics and b is a constant chosen to reflect the variability of symmetry equivalent strong reflections.

Crystallographers commonly use 'weighted R ', wR , to measure the goodness of fit of a refinement solution $\hat{\theta}$. Specifically,

$$wR = \left[\sum_{i=1}^N r_i^2(\hat{\theta}) / \sum_{i=1}^N w_i y_i^2 \right]^{1/2} \quad (4)$$

Weighted R has the same numerator as the residual standard deviation estimate of scale,

$$SD = \left[\sum_{i=1}^N r_i^2(\hat{\theta}) / (N - p) \right]^{1/2} \quad (5)$$

Suppose that the weight used in the refinement is obtained by neglecting counting statistics in (3), giving a 'constant relative variance weight'. Suppose further that this weight correctly assays precision except, possibly, that the relative variance factor b^2 may be wrong. Let b_0^2 be the correct factor. Then for large N wR estimates $\sqrt{1 - p/N} b_0$ while SD estimates b_0/b . Thus, for situations where relative variance error in (3) is dominant and b is at least approximately correct wR estimates relative precision and SD estimates unity.

In a full-matrix refinement, multiple equivalent reflections and duplicate reflections are often handled by first averaging to get a single value and then modifying the weight to reflect the precision of a weighted average where individual weights reflect the relative precision of the individuals making up that average. For simplicity, in the remainder of the discussion, the term 'equivalent' is used both for a true equivalent reflection, where crystal symmetry gives the same model for several distinct incident beam directions, and for duplicate reflections where repeated measurements are made for the same beam orientation.

When a robust/resistant algorithm is used to perform the refinement, all multiple equivalent reflections should be included as individual observations. This allows the fitting algorithm to downweight individual reflections which are discrepant. Now the measures of agreement, wR and SD , should calculate goodness of fit using the difference between the weighted average of equivalent reflections and the estimated model. That is, the variability of equivalent

reflections about their weighted average should be removed from those measures.

Consider a set of N_i equivalent reflections y_{ij} where $j = 1, 2, \dots, N_i$. The fitted model is $m_i(\hat{\theta})$. The contribution of these reflections to the weighted sum of squares of residuals can be partitioned formally as follows

$$\sum_j r_{ij}^2(\hat{\theta}) = w_i (\bar{y}_i - m_i(\hat{\theta}))^2 + \sum_j w_{ij} (y_{ij} - \bar{y}_i)^2, \quad (6)$$

where

$$w_i = \sum_{j=1}^{N_i} w_{ij} \text{ and } \bar{y}_i = \sum_{j=1}^{N_i} w_{ij} y_{ij} / w_i.$$

The first term in (6) is the measure of agreement and the second term is the variability of the individual observations about their weighted average. There is no analogous partition formula for the denominator of wR . However, wR type functions measuring the agreement of the weighted average of equivalent reflections and the internal variability among reflections can be calculated and compared with the usual agreement measure using all reflections. The specific formulas for wR and SD are listed in the top two rows of Table 1. Discussion of the robust/resistant agreement measure SH in Table 1 is deferred until all needed notation is defined.

As remarked above, weighted least-squares refinement can give very poor estimates if the error structure is not Gaussian. Sets of structure factors are not uniform in precision. Weak reflections are likely to have major errors. Instrumentation effects can introduce orientation and absorption biases not included in the model $m_i(\theta)$. Crystallographers take care of discrepant observations by screening data prior to the full matrix refinement calculation. Computer programs (for example, Finger and Prince, 1975) include quality control steps for the rejection of

Table 1. Measures of agreement with multiple equivalent and/or duplicate reflections

	Agreement		Internal variability
	1)	2)	
$(wR)^2$	$\sum_{i=1}^N w_i [\bar{y}_i - m_i(\hat{\theta})]^2$	$\frac{N}{\sum_{i=1}^N w_i \bar{y}_i^2}$	$\frac{\sum_{i=1}^N \sum_{j=1}^{N_i} w_{ij} (y_{ij} - \bar{y}_i)^2}{\sum_{i=1}^N \sum_{j=1}^{N_i} w_{ij} y_{ij}^2}$
$(SD)^2$	$\sum_{i=1}^N w_i [\bar{y}_i - m_i(\hat{\theta})]^2$	$\frac{N}{N-p}$	$\frac{\sum_{i=1}^N \sum_{j=1}^{N_i} w_{ij} (y_{ij} - \bar{y}_i)^2}{\sum_{i=1}^N (N_i - 1)}$
$(SH)^2$	$\sum_{i=1}^N \frac{\sum_{j=1}^{N_i} \{\phi[r_{ij}(\hat{\theta}/s)]\}^2 w_{ij} [\tilde{y}_i - m_i(\hat{\theta})]^2}{\beta (N-p)}$		$\frac{\sum_{i=1}^N \sum_{j=1}^{N_i} \{\phi[r_{ij}(\hat{\theta}/s)]\}^2 w_{ij} (y_{ij} - \bar{y}_i)^2}{\beta \sum_{i=1}^N (N_i - 1)}$

- 1) For weighted least-squares refinements \bar{y}_i is the w_{ij} weighted average of all equivalent reflections, thus $\bar{y}_i = \sum_j w_{ij} y_{ij} / w_i$ and $w_i = \sum_j w_{ij}$.
- 2) For biweight robust/resistant refinements \tilde{y}_i is a $\phi_{ij} w_{ij}$ weighted average over all equivalent reflections, thus $\tilde{y}_i = \sum_j \phi[r_{ij}(\hat{\theta})/s] w_{ij} y_{ij} / \sum_j \phi[r_{ij}(\hat{\theta})/s] w_{ij}$.

structure factors which do not fit the model. However, there are problems with this crystallographer controlled screening approach. Fitting by non-linear weighted least squares to a model with hundreds of parameters is very complex. Outlier data can only be screened if residuals are excessive. In such a complex fitting problem the weighted least-squared algorithm makes minor adjustments in parameters to fit discrepant data. Such data may not be detectable by a residual test.

Most weighted least-squares refinement computer programs can easily be modified to be more robust/resistant. Both weighted least squares and the modification are examples of a class of estimation methods which, for crystal structure refinement, take the form, minimize the loss function

$$f(\theta) = \sum_{i=1}^N \rho [r_i(\theta)/s], \quad (7)$$

by selecting θ so that the gradient, ∇f , vanishes. Here s is a resistant estimate of measurement error scale or uncertainty based on residuals. The defining equations for the vanishing of ∇f (except for a constant) can be written as

$$\frac{\partial f(\theta)}{\partial \theta_j} = \sum_{i=1}^N \phi [r_i(\theta)/s] w_i^{1/2} r_i(\theta) \frac{\partial m_i(\theta)}{\partial \theta_j} = 0, \quad j = 1, 2, \dots, p, \quad (8)$$

where $\phi(x) = (1/x) d\rho(x)/dx$. Weighted least squares is the special case $\rho(x) = x^2/2$. Then, $\phi(x) = 1$ and the effect of the residual $r_i(\theta)$ on the solution to (8) is proportional to residual magnitude—an unresistant situation.

A robust/resistant alternative to weighted least squares has $\phi(x)$ near to 1 for x close to zero, decreasing toward zero or possibly exactly zero for $|x|$ large, and a smooth transition in between. Andrews (1974) and Beaton & Tukey (1974) illustrate the

robust/resistant approach for simple linear regression situations. The specific estimates are not highly dependent on the exact shape of $\phi(x)$. Here we use the Tukey (1974) 'biweight' function

$$\begin{aligned}\phi(x) &= [1 - (x/a)^2]^2 \quad \text{for } |x| \leq a; \\ &= 0 \quad \text{otherwise.}\end{aligned}\quad (9)$$

This corresponds to a loss function

$$\begin{aligned}\rho(x) &= (x^2/6) [1 + \phi^{1/2}(x) + \phi(x)], \quad \text{for } |x| \leq a; \\ &= a^2/6 \quad \text{, for } |x| > a.\end{aligned}\quad (10)$$

The shape of this function is shown in Fig. 1 and compared with the parabola $\rho(x) = x^2/2$ and also with the function $\rho(x) = c \ln [1 + (x/b)^2]$, which is a

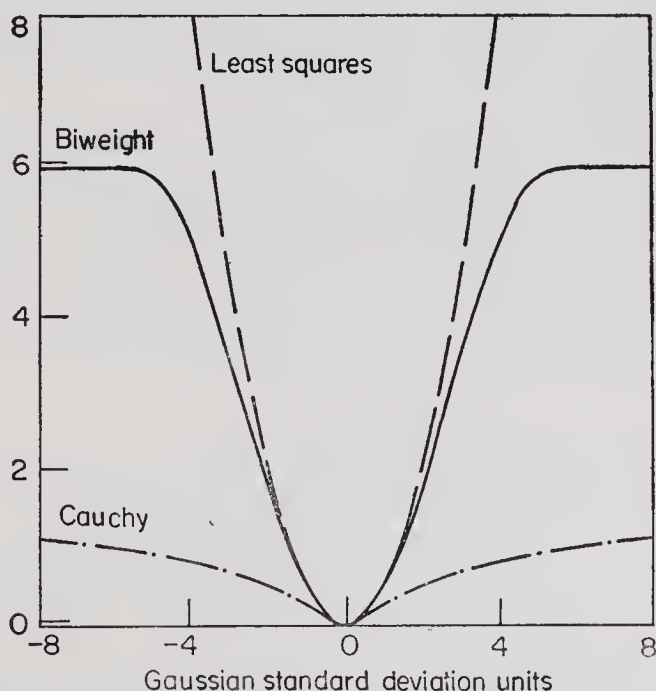


Fig. 1. Least squares compared to two alternative loss functions.

loss function that would lead to maximum likelihood estimates if the errors had a Cauchy distribution. The function of (10) lies close to the least-squares function for small x , but is constant for large x , so that large deviations do not influence the solution of (8). (Because the function $\phi(x)$ appearing in the algorithm for finding the minimum is a factor multiplying the weight, this technique is sometimes called 'iteratively reweighted least squares'. While this term is suggestive, it should be clearly understood that the function being minimized is quite distinct from a sum of squares.)

For comparative purposes in Fig. 1, $b = 0.6745$ so the Gaussian and Cauchy error distributions have the same *probable* error. Also, $c = b^2/2$ so the three functions have unit curvature at the origin and, hence, weight reflections with small residuals identically. Based on probable error normalization the biweight type loss function (10) places much more weight on large residuals than does the Cauchy-type function. This is reasonable since error distributions as long-tailed as a Cauchy distribution are rarely, if ever, encountered in real experimental data.

The crystal structure model $m_i(\theta)$ is a non-linear function of θ . With $\phi(x) = 1$ in (8), iterative approaches have been used to obtain the solution. Busing, Martin & Levy (1962) and Finger and Prince (1975) describe algorithms and computer programs for iterative solution of the weighted least-squares equations. In general, such algorithms include:

- (a) A linearized form of (8) with $\phi(x) = 1$ for calculating θ^{q+1} from a previous solution $\hat{\theta}^q$;
- (b) A stopping rule for deciding when the iterative procedure has converged to an acceptable solution $\hat{\theta}$; and
- (c) A procedure for calculating the precision of the estimate $\hat{\theta}$.

Newton linearization of (8) gives

$$\Delta\theta_j^{q+1} = \sum_{k=1}^p C^{jk} \sum_{i=1}^N \phi[r_i(\theta^q)/s^q] w_i^{1/2} r_i(\theta^q) \times \frac{\partial m_i(\theta^q)}{\partial \theta_k}, \quad (11)$$

where C^{jk} is the generic element from the inverse of the Hessian matrix of $f(\theta)$. Specifically,

$$C_{jk} = \sum_{i=1}^N w_i \omega[r_i(\theta^q)/s^q] \frac{\partial m_i(\theta^q)}{\partial \theta_j} \frac{\partial m_i(\theta^q)}{\partial \theta_k} - \sum_{i=1}^N w_i^{1/2} \phi[r_i(\theta^q)/s^q] r_i(\theta^q) \frac{\partial^2 m_i(\theta^q)}{\partial \theta_j \partial \theta_k}. \quad (12)$$

In (12) $\omega(x) = (d/dx)(x\phi(x)) = (d^2/dx^2)\rho(x)$. In crystallography, it is customary to replace the structure factor expression (1) by a linear Taylor expansion. Hence, the second partial term does not appear in (12). Following Huber (1973), we simplify the first sum by replacing each ω factor by the average over all N observations. The result is the simplified version

$$C_{jk} = \bar{\omega} \sum_{i=1}^N w_i \frac{\partial m_i(\theta^q)}{\partial \theta_j} \frac{\partial m_i(\theta^q)}{\partial \theta_k}, \quad (13)$$

where

$$\bar{\omega} = (1/N) \sum_{i=1}^N \omega[r_i(\theta^q)/s^q].$$

In (13), $\bar{\omega}$ plays the role of a variance efficiency factor with respect to Gaussian error structure. That is, the parameter estimates have variances of the same order as they would have if the error distribution were Gaussian and there were $\bar{\omega}N$ reflections in the data set.

In (11), the width parameter s^q is a resistant estimate constructed from residuals. Andrews (1974), Tukey

(1974), and Welsch & Kuh (1977) are examples of those suggesting $s^{q+1} = \text{MAD}^q/0.675$ where MAD^q is the Median Absolute Deviation; *i.e.* the median of the absolute residuals, $|r_i(\theta^q)|$.

Huber (1973) suggests $s^{q+1} = (\alpha^q/\beta)^{1/2}$ as a resistant estimate of scale. Here,

$$\alpha^q = \sum_{i=1}^N \{\phi[r_i(\theta^q)/s^q]\}^2 r_i^2(\theta^q)/(N-p), \quad (14)$$

and β is the expected value of $[Z\phi(Z)]^2$ with Z distributed according to the true error law. If the error law is Gaussian and the biweight function (9) cut-off is $a=6$, $\alpha\beta=0.72767$ makes α^q/β an unbiased estimate of the variance of the distribution. For a longer tailed error law, β is smaller, but the practical application of the method is not highly dependent on the choice of β . With multiple equivalent reflections, (14) must be modified in a manner similar to that for weighted least squares. Calculationally, in determining the solution $\hat{\theta}$, the product term ϕw is just a least squares type weight. In the normal equation (8), the terms involving an equivalent set of reflections y_{ij} ($j=1,2,\dots,N_i$) can be rewritten as

$$\begin{aligned} & \sum_{j=1}^{N_i} \phi[r_{ij}(\theta)/s] w_{ij}^{1/2} r_{ij}(\theta) \frac{\partial m_i(\theta)}{\partial \theta_a}, \quad a=1,2,\dots,p \\ &= \left\{ \sum_{j=1}^N \phi[r_{ij}(\theta)/s] w_{ij} \right\} [\bar{y}_i - m_i(\theta)] \frac{\partial m_i(\theta)}{\partial \theta_a}, \quad (15) \\ &= \bar{\phi}_i w_i^{1/2} \bar{r}_i(\theta) \frac{\partial m_i(\theta)}{\partial \theta_a}, \end{aligned}$$

where

$$\bar{\phi}_i = \sum_{j=1}^{N_i} \{ \phi[r_{ij}(\theta)/s] \}^2 w_{ij} / \sum_{j=1}^{N_i} \phi[r_{ij}(\theta)/s] w_{ij};$$

$$\bar{r}_i(\theta) = w_i^{1/2} [\bar{y}_i - m_i(\theta)];$$

$$w_i = \sum_{j=1}^{N_i} \{ \phi[r_{ij}(\theta)/s] w_{ij} \}^2 / \sum_{j=1}^{N_i} \{ \phi[r_{ij}(\theta)/s] \}^2 w_{ij};$$

and

$$\bar{y}_i = \sum_{j=1}^{N_i} \phi[r_{ij}(\theta)/s] w_{ij} y_{ij} / \sum_{j=1}^{N_i} \phi[r_{ij}(\theta)/s] w_{ij}.$$

Thus, as for the usual weighted least squares calculation with duplicate observations, a structure factor robust/resistant fit depends upon an equivalent set of reflections only through a weighted average \bar{y}_i . Assuming that $1/w_{ij}$ is the variance of y_{ij} , then $1/w_i$ of (15) is the variance of \bar{y}_i . Now $\bar{\phi}_i$ plays the role of the biweight for $w_i^{1/2} \bar{y}_i$. The formula (15) has the same form as the normal equation formula (8), *i.e.* a quadruple product of a weight adjustment factor, the reciprocal of a standard deviation, a residual, and a partial derivative. Applying an asymptotic argument due to Huber (1973), the contribution to the measure of agreement (14) simplifies to

$$\phi_i^2 [\bar{r}_i(\theta^q)/s^q]^2 = \left\{ \sum_{j=1}^{N_i} \{ \phi[r_{ij}(\theta^q)/s^q] \}^2 w_{ij} \right\} [\bar{y}_i - m_i(\theta^q)]^2. \quad (16)$$

Summation over i in (16) gives the agreement term in row (SH)² of Table 1. The internal variability term applies Huber's argument to estimating duplicate error.

Equation (11), with s^{q+1} defined by MAD or by (14), gives an iterative procedure for solving the system (8).

A computer algorithm for solving the weighted least-squares refinement problem can be changed into a robust/resistant algorithm by the inclusion of $\phi[r_i(\theta^q)/s^q]$ in (11), the inclusion of $\bar{\omega}$ in (13), and an extra equation defining s^q . For the biweight function (9), $\omega(x) = 5\phi(x) - 4\phi^{1/2}(x)$. Thus $\bar{\omega}$ is easily calculated from $\phi[r(\theta^q)/s^q]$ values. Our experience suggests that, with either of the above procedures for estimating width, a value of the constant a in (9) of about 6 screens out extreme outliers while having a minor effect on the body of the data.

For best results, a robust/resistant iterative regression calculation should start from a resistant estimate of θ . For simple linear regression situations, Andrews (1974) suggests a median procedure for getting the initial θ' estimate. The complexity of the crystal structure model suggests that a more practical approach is to start with whatever estimate is available, and accept a penalty of more iterations. The stopping rule should not be based on a fixed number of iterations. We have observed that for some data sets convergence is much faster than with a conventional weighted least-squares algorithm and for some much slower. We use a stopping rule of nominal maximum parameter adjustment measured in standard deviation units.

A number of approaches are available for estimating the standard deviations of parameter estimates, $\hat{\theta}$, based on asymptotic expansion arguments; *i.e.* N , p and $N-p$ must all be large. Some approaches (see Mallows (1973) and Welch (1975)) estimate the covariance of $\hat{\theta}$ with a scalar multiple of the inverse of a matrix with jk -th element

$$C'_{jk} = \sum_{i=1}^N \phi[r_i(\theta)/s] w_i \frac{\partial m_i(\hat{\theta})}{\partial \theta_j} \frac{\partial m_i(\hat{\theta})}{\partial \theta_k}.$$

These approaches do the estimation by using weights $\phi[r_i(\hat{\theta})/s] w_i$ in a standard weighted least-squares

program. Here, the dependence of $\phi[r_i(\hat{\theta})/s]$ on $\hat{\theta}$ is ignored in the expansion (11). The basic assumption is that $1/\phi[r_i(\hat{\theta})/s]w_i$, not $1/w_i$, is proportional to the variance of y_i . Our approach assumes that all the weighted structure factor amplitudes, $w_i^{1/2} y_i$, have unit variance. They are random observations from a long tailed error distribution, so that there is high probability that a few structure factors will have extreme errors. We use $\phi[r_i(\hat{\theta}/s)]$ as a calculational convenience to reduce the influence of the extreme data. Thus, we follow Huber (1973) and use a scalar multiple of C^{-1} , where $C = \{C_{jk}\}$ is defined by (13), to estimate the covariance matrix of $\hat{\theta}$. Specifically, our variance estimate of $\hat{\theta}$ is

$$S_{\hat{\theta}_j}^2 = K\beta(SH)^2 C^{jj}. \quad (17)$$

Here, $(SH)^2$ is defined in Table 1. C^{jj} is the j -th diagonal element of C^{-1} . K is a *bias correction factor* defined as

$$K = 1 + p(1 - \bar{\omega})/N\bar{\omega}. \quad (18)$$

Huber's development of (17) assumes that the error distribution and the loss function ρ are symmetric, and that all the diagonal elements of the C matrix (C_{ii} of (13)) are identical. The first two assumptions are reasonable for crystal structure refinement. The last is never satisfied. Hence (17) is an approximation that needs empirical verification.

Application to D(+)-tartaric acid

The Single-Crystal Intensity Measurement Project sponsored by the Commission on Crystallographic Apparatus of the International Union of Crystallography (Abrahams, Hamilton, & Mathieson, 1970, hereinafter referred to as AHM) was aimed at deter-

mining the level of consistency that could be obtained in the collection of *X*-ray diffraction intensity data from single crystals of one compound in different laboratories. Crystals of *D*(+)-tartaric acid were grown in one laboratory and distributed to 16 laboratories in eight countries. These laboratories in turn collected data sets from the crystals they received, using their established techniques. The data sets were then returned to a single laboratory and compared by statistical methods, and the analysis showed that there were some fairly substantial differences among the various data sets.

In an attempt to determine what effect these differences would have on the results of a structure refinement, Hamilton & Abrahams (1970, hereinafter referred to as HA) refined the structure of *D*(+)-tartaric acid using 10 of the 17 data sets. (One laboratory submitted two data sets. Three sets had insufficient data to refine, and in four the least-squares procedure failed to converge.) Again, there were some substantial differences, from one data set to another, in the parameter estimates obtained from the refinement procedure. Subsequently, Mackenzie (1974) examined the nature of systematic differences among the various data sets. He concluded that there was a tendency for the differences to be greatest for the largest magnitude structure factors, suggesting that an important difference among the experiments was the amount of secondary extinction present in the particular crystal used in each experiment.

In our study we have applied the robust/resistant method described in the previous section to 12 of the Single-Crystal Intensity Measurement Project data sets in order to determine if the new approach would reveal more about the nature of the biases in the experiments and resolve some of the discrepancies in the results. In the following discussion, we use the HA numbers to identify experiments. The experiments selected for reanalysis include 2, 4, 5, 7, 8, 9, 11a, 12, 13, 14, 15 and 16. Our twelve include the ten refined by

HA plus experiments 12 and 14, for which their refinements diverged. The full set of structure factors consist of every hkO reflection, including equivalent reflections, within the range $(\sin \theta)/\lambda \leq 0.5 \text{ \AA}^{-1}$, and all reflections with positive k and l within the same $(\sin \theta)/\lambda$ range, except in experiment 14 where $\bar{h}kl$ reflections were replaced by hkl equivalent ones. The full set includes 332 non-equivalent reflections. Only experiments 2, 14 and 16 measured all 332. We included in our analysis all the experiments for which at least 232 non-equivalent reflections were measured.

The refinement for each data set had three stages: (1) an attempt to 'recreate' the results of HA by repeating as closely as we could the conditions of their refinements; (2) a refinement including secondary extinction; and finally (3) a refinement using the robust/resistant procedure described above.

The $D(+)$ -tartaric acid structure belongs to space group $P2_1$, and the unit cell contains 32 atoms—12 oxygen, 8 carbons, and 12 hydrogens. Therefore the positions and thermal parameters for 16 atoms must be refined in the complete model. (Any single y -coordinate must be fixed to define the origin.)

The computer program RFINE4 (Finger and Prince, 1975) was modified as described in the previous section to perform the robust/resistant fitting of the model (1) using the Tukey biweight function (9). In the complete model, including extinction, there are 115 parameters to be refined. These parameters include 47 position parameters, 66 temperature parameters, a scale factor, and an extinction parameter. The extinction parameter, r^* (Zachariasen, 1968) has the form

$$F'_{\text{calc}} = F_{\text{calc}} [1 + \beta(\theta) F_{\text{calc}}^2 r^*]^{-1/4}; \quad (19)$$

where

$$\beta(\theta) = \{ [2e^4 \lambda^2 (1 + \cos^4 2\theta)] / [m^2 c^4 V^2 (1 + \cos^2 2\theta)] \} \bar{T}.$$

λ is the wavelength, V is the unit cell volume, e and m are the charge and mass of the electron, and c is the

velocity of light. In the absence of absorption information for determining the pathlength parameter \bar{T} , this quantity was treated as a constant, and the refined extinction parameter corresponds to the product $\bar{T}r^*$.

In the model (1), the scattering factors f_j are determined in the manner of Cromer and Mann (1968). RFINE4 uses an analytic function with coefficients chosen to fit smooth curves through the tabulated points (International Tables for X-ray Crystallography, 1962).

The constant relative variance weight (3) with $\sigma_{st}^2=0$ and $b=0.1$ was used by HA in their refinements. Since we wished to reproduce their results as closely as possible, the same weight function was used. A more general weight function of the form (3) with $\sigma_{st}^2>0$ would improve the quality of the refinements by a better approximation of the random errors in weak reflections. However, since all our refinements use the same weights, we have a valid, though not necessarily optimum, comparison of classical and robust/resistant approaches to crystal structure refinement.

In the first refinement of each experiment we used the same initial parameters for position and thermal vibration as did HA. The heavy atom parameters were those of Okaya, Stemple & Kay (1966). The hydrogen atom parameters were preliminary neutron diffraction results of Cox, Sabine & Taylor (1966). For each experimental data set, preliminary iterations were done by refining only the scale factor. This was followed by iterations with the scale factor and the heavy atom parameters being refined until convergence. Finally, iteration with refinement of all parameters was continued until convergence. In each case, convergence was defined as maximum $|\Delta\theta|/S(\theta) \leq 0.10$. To see whether the starting point would influence the solution, some refinements were repeated beginning with the final results of HA. In all cases, the two refinements converged to the same solution, which differed slightly, for reasons that are not clear, from

those of HA. In our refinements to recreate the HA results, we used all of the data reported by each experimenter which satisfied $\sin \theta/\lambda < 0.5 \text{ \AA}$. The only screening was by RFINE4, which rejects discrepant reflections based on $|r_i(\theta)| > \text{constant}$. (In our case equal to 2.) Table 2 classifies the sets of reflections for each experiment into non-equivalent, additional equivalent, and total number of reflections. The last column is the number of reflections used by HA in their refinements. Experiment 5 included three reflections, 100, 002, and 020, which were duplicated 37 times each. We used two measurements of each reflection in our refinement.

The results of the recreated refinement were used as the initial parameter estimates in the second refinement. Experiment 11a did not refine to convergence, as several hydrogen atom parameters continued to oscillate. The iteration with smallest SD was used to start the extinction refinement. Initially, refinement was restricted to scale and extinction factors. After reasonable convergence, iteration continued with refinement of the full model. All experiments except

Table 2. *Structure factor sets satisfying IUC guides and Hamilton and Abrahams Total*

Experiment Number	Non-Equivalent	Equivalent	Total	HA
2	332	35	367	368
4	331	0	331	331
5	327	178	605	607
7	331	101	432	429
8	320	35	355	355
9	303	65	368	366
11a	251	88	339	342
12	303	99	402	403
13	326	92	418	417
14	332	0	332	273
15	234	88	322	324
16	332	96	428	429

11a and 12 converged. Experiment 11a diverged, while in experiment 12 some hydrogen parameter estimates oscillated.

The results of the extinction refinement were used as initial parameters in the robust/resistant refinements. For all experiments, (except experiment (5) at least ten iterations were needed to bring solid convergence. As the iterations progressed, there was a gradual change in the residual configuration, a tightening of the majority of the residuals, and a migration away from the body of the data by a small fraction. The lack-of-convergence problem evident in earlier refinements did not carry over to the robust/resistant refinement.

Discussion of results

RFINE4 constructs the usual standard deviation estimates based on the Hessian matrix of second partial derivatives and the residual standard deviation estimate (5). Program output includes the change from iteration to iteration of all parameters, measured in standard deviation units. In all cases the final stopping rule was that the maximum change in parameters was not more than 10% of the standard deviation. Clearly, this rule is conservative, particularly for robust/resistant refinements where standard deviations are in general underestimated by a least squares type algorithm.

Table 3 summarizes the data set sizes and scale measure comparisons between our recreated refinements and the HA refinements. For the recreated refinements, the formulas in Table 1 were used to calculate weighted $R(wR)$ and standard deviation (SD) estimates of scale. The agreement columns (A) are measures of fit between the full 115 parameter model (1) and the data set of non-equivalent reflections (some being weighted averages of the equivalent reflections for a single triple of Miller indices). The

Table 3. Comparison of data set size and scale measures for recreated and Hamilton and Abrahams refinements

Exp. No.	SD				RECREATED				wR				HA	
	A	D.F.	IV	D.F.	T	A	IV	T	N	T	N	T	N	wR
2	0.61	202	0.14	33	0.57	0.046	0.010	0.047	352	0.091	368	0.091	368	
4	0.78	198	—	0	0.78	0.062	—	0.062	316	0.107	331	0.107	331	
5	0.49	210	0.42	175	0.46	0.032	0.037	0.040	505	0.043	607	0.043	607	
7	0.53	203	0.22	100	0.45	0.037	0.018	0.039	420	0.052	429	0.052	429	
8	0.59	193	0.39	36	0.56	0.044	0.027	0.046	344	0.093	355	0.093	355	
9	0.54	178	0.26	66	0.48	0.038	0.021	0.039	360	0.066	366	0.066	366	
11a	0.73	129 ¹⁾	0.63	85	0.68	0.046	0.053	0.055	331	0.123	342	0.123	342	
12	0.71	169	0.31	89	0.60	0.048	0.026	0.050	371	²⁾	403	²⁾	403	
13	0.57	188	0.33	92	0.50	0.039	0.028	0.042	396	0.060	417	0.060	417	
14	0.63	188	—	0	0.63	0.050	—	0.050	301	²⁾	273	²⁾	273	
15	0.48	114	0.17	87	0.38	0.029	0.014	0.030	316	0.039	324	0.039	324	
16	0.46	205	0.29	92	0.42	0.033	0.024	0.035	413	0.043	429	0.043	429	

¹⁾ Oscillatory behaviour. Data for minimum total SD value reported.²⁾ Diverged.

size of the data set involved in the fit is $115+$ degrees of freedom. After outlier screening, the smallest data set for fitting is the 229 of experiment 15.

The internal variability columns (IV) are measures of scale computed from the differences between equivalent reflections totalled over all sets of equivalent reflections. For experiments 4 and 14 only averages of equivalent reflections were reported. One degree of freedom is lost for each set, so the degrees of freedom (DF) column is the total number of equivalent reflections minus the number of distinct sets. In all cases, the standard deviation (SD) for internal variability is smaller than that for agreement. In all but two cases, the weighted $R(wR)$ for internal variability is smaller than that for agreement. This is very reasonable since internal variability measures duplication error and experimental asymmetry with respect to equivalent beam positions, while, in addition, agreement measures inadequacies of the crystal structure model.

Among all experiments (excluding 11a, which did not converge) experiment 5 stands out as having very little variability above that suggested by equivalent reflections. That is, only for experiment 5 does the model appear to fit the data adequately relative to the scatter of equivalent reflections.

The SD total column (T) is the average of agreement and internal variability SD values weighted according to the respective numbers of degrees of freedom. This is often called the 'estimated standard deviation of an observation of unit weight' by crystallographers. It must be intermediate between the agreement SD and the internal variability SD. If the agreement SD is larger, the total SD is always an underestimate of discrepancy between data and model and, hence, exaggerates the precision of the structure factor data. There is no such order relationship among the wR values because the total is not a weighted average. HA do not include SD values, but only present totals R and wR . In Table 3, we compare the

two refinements. The total number of reflections in the data set is the sum of agreement degrees of freedom, internal variability degrees of freedom, and the number of parameters, 115. In all cases, our recreated wR values are smaller, sometimes by as much as a factor of 2. In all our refinements, we screened out extreme reflections with $|r_i| > 2$. Apparently HA did not screen the data as extensively. Their data sets (excluding experiment 14) are larger. We cannot say whether our refinement is generally a better fit, with uniformly smaller standardized residuals or whether the inclusion of a few discrepant reflections has artificially increased the HA wR values.

Table 4 lists wR and SD measures of agreement for our three refinements, recreation, extinction, and biweight. The tabulated values are number (N), weighted $R(wR)$, and standard deviation (SD), calculated from non-equivalent reflections by the formulas in Table 1. In addition, for biweight, the Huber standard deviation (SH) and the average effectiveness of an observation ($\bar{\omega}$) are listed. The three N , wR and SD columns for the recreated refinement are reproduced from Table 3 for easy comparison.

We first note that there was one experiment, 11a, for which the recreated refinement did not converge. The refinement reached a point of oscillatory behaviour, with little change in SD, but swings, from iteration to iteration, of several standard deviation units for two of the thermal vibration hydrogen atom parameters. For the extinction refinement experiments 11a and 12 both diverged. A smallest standard deviation of agreement (SD) was reached for which there were still significant shifts in the parameters. Then, with additional iterations SD diverged. For six experiments the extinction refinement used more observations than the recreated ones. A closer look at individual structure factor data shows that the strong reflections, with presumably large amounts of extinction, were brought in line with the body of the data and, hence,

Table 4. *Agreement scale measures*

Exp. No.	Recreated			Extinction			Robust/Resistant				$\bar{\omega}$
	N	wR	SD	N	wR	SD	N	wR	SD	SH	
2	317	0.046	0.61	330	0.032	0.41	329	0.027	0.34	0.38	0.885
4	313	0.062	0.78	325	0.050	0.62	322	0.041	0.50	0.56	0.881
5	325	0.032	0.49	325	0.030	0.47	325	0.029	0.43	0.48	0.861
7	318	0.037	0.53	331	0.034	0.48	320	0.022	0.30	0.32	0.866
8	308	0.044	0.59	314	0.041	0.49	311	0.029	0.38	0.42	0.820
9	293	0.038	0.54	295	0.038	0.53	293	0.030	0.42	0.46	0.859
11a	—	—	—	—	—	—	248	0.049	0.76	0.85	0.902
12	284	0.048	0.71	—	—	—	286	0.041	0.60	0.66	0.833
13	303	0.039	0.57	303	0.038	0.55	306	0.035	0.50	0.55	0.846
14	303	0.049	0.63	316	0.032	0.40	316	0.029	0.35	0.39	0.826
15	228	0.027	0.46	228	0.027	0.46	228	0.026	0.43	0.48	0.887
16	320	0.033	0.46	320	0.025	0.36	317	0.021	0.29	0.32	0.853

satisfied the cut-off rule of $|r_i| < 2$. For the robust/resistant refinements all the twelve experiments converged solidly. The maximum parameter shift in the final iteration, measured in standard deviation units, was of the order of a percent or two. Thus, not only did the robust/resistant algorithm force convergence for the two experiments which previously did not converge, it also brought solid convergence for other experiments which were still having parameter shifts of 5 to 10% of their standard deviations.

The number of observations column (N) for the robust/resistant refinements gives the number of reflections which received positive weight. Since the biweight function decreases monotonically to zero, there are a number of these reflections which receive very low weight. The average effectiveness column ($\bar{\omega}$) suggests that the refinements have the precision of ones with about 15% of the reflections eliminated. Among those that refined to convergence the experiments with major extinction corrections are 2, 4, 7, 8 and 14. These experiments have the largest $\bar{\text{Tr}}^*$ extinction parameter estimates, use more reflections in the extinction refinement than in the recreated refinement, and have a significant reduction in the standard deviation of agreement. Mackenzie's (1974) raw data residual analysis clearly picks out experiments 2 and 4 as having major extinction, but the status of the other three is unclear. Thus the inclusion of the extinction factor for these experiments brings the strongly extinguished reflections in line with the model, allows them to be included in the refinement and, at the same time reduces the standard deviation of agreement.

There is a significant positive correlation between the degree of extinction, as measured by the fitted parameter $\bar{\text{Tr}}^*$ of (19), and the crystal volume reported by experimenters in AHM. This is what would be expected because the values of $\bar{\text{T}}$ are larger in larger crystals. For the experiments without strong extinc-

tion there is very little difference between the standard deviations of agreement for recreated and extinction refinements, although in no case is the extinction one larger. There is no strong pattern between the standard deviations of agreement for the extinction refinements and the Huber values for the robust/resistant refinements. In general, the Huber values are slightly smaller. The standard deviation of agreement values obtained from the refinement are clearly biased low, as pointed out above. The Huber values are approximately 10% larger and appear to compensate for most of that bias, but still may be slightly low. The underlying error distribution is not Gaussian, so the β value in the Huber variance formula is too large.

If there are significant biases in the individual parameter estimates, then the standard deviation values for these parameter estimates do not account for the differences among the parameters across experiments. Hence in a manner similar to HA we selected a model group of experiments which are free of obvious parameter bias. In order to do that we looked at two sets of parameters across the extinction experiments—the atom position parameters and the heavy atom diagonal thermal parameters. For each parameter in the model, the twelve parameter estimates were ranked from smallest to largest, to determine whether any of the experiments showed up consistently as having the extreme parameters. For heavy atom position parameters in experiment 15, six of the ten *x*-coordinate position parameters were the highest. In addition two *y*-coordinate position parameters were the highest and five were the lowest. Thus one experiment accounted for 13 out of the 19 extreme *x* and *y* coordinate position parameter estimates.

Using a similar ranking on the heavy atom diagonal thermal parameters, experiments 11a, 12 and 13 account for all but one of the largest parameter estimates and 17 out of the 29 smallest parameter estimates. Experiment 13 ZZ-coordinate diagonal thermal parameters are 2 to 5 times larger than

those for the rest of the experiments. Experiment 12 has 8 of the largest XX-diagonal thermal parameters and 9 of the smallest ZZ-diagonal thermal parameters. Experiment 11a has 8 of the smallest YY-diagonal thermal parameters and 9 of the YY-largest diagonal thermal parameters. Based on the result of this simple extreme value screening of the heavy atom parameters, experiments 11a, 12, 13 and 15 were eliminated from further consideration. Experiments 11a and 12 could just as well have been eliminated because of convergence problems. Also, experiment 15 clearly has an insufficient number of structure factor values (228) to estimate anisotropic thermal vibration parameters.

HA found that the standard deviations of parameter estimates were too small to predict the spread among parameters across the individual experiments. The situation was particularly bad for thermal vibration parameters, where the standard deviations were small by a factor of about four. They attributed this to the fact that some of the experiments had serious systematic errors, and, hence, the model was incorrect. Since inclusion of extinction removes one source of systematic error, one might expect that the standard deviations would come closer to predicting the total scatter in parameter estimates. A further question was whether the robust/resistant approach would compare favourably with the classical extinction model, or possibly improve the situation by down weighting a small fraction of the structure factor estimates which had additional serious systematic errors. For the recreated and extinction refinements, we estimate the standard deviations of individual parameter estimates by changing the multiplier on diagonal elements of the inverse of the Hessian matrix from the standard deviation of unit weight calculated from all of the residuals to the standard deviation of agreement as given in Table 3. For the robust/resistance refinements, the standard deviations of parameter estimates are calculated from (17). Using these standard deviation

estimates, a chi-square parameter agreement statistic was calculated across the eight 'good' experiments using the formula.

$$\chi_a^2 = \sum_{i=1}^8 \left(\frac{P_{ai} - \bar{P}_a}{SD_{ai}} \right)^2, \text{ where}$$

$$\bar{P}_a = \left(\sum_{i=1}^8 P_{ai} / \overline{SD_{ai}^2} \right) / \left(\sum_{i=1}^8 1 / \overline{SD_{ai}^2} \right).$$

Here χ_a^2 is the agreement statistic for the a -th parameter, and \bar{P}_a is the weighted (by the reciprocal of the estimate of variance) estimate of the mean of the a -th parameter for the eight 'good' experiments. If the total variability across the parameter set is explained by the refinement standard deviations, this agreement statistic is approximately distributed as a chi-square variable with seven degrees of freedom. If it is significantly large, then either the precision of some of the estimates is less than that indicated by the refinement standard deviations, or there are systematic effects in some of the experiments which make their parameter estimates outliers.

Table 5 is a stem and leaf display of the chi-square agreement statistics for the 113 individual parameters (excluding scale and extinction) in the complete 16 atom model (1). For ease of comparison across the three methods of analysis (recreated, extinction and robust/resistant) the chi-square values have been symmetrized using a logarithmic transformation. Specifically, if χ^2 is a chi-square variable on k degrees of freedom, then $Y = (k/2)^{1/2} \ln(\chi^2/k)$ is nearly symmetrically distributed, with mean and variance near 0 and 1 respectively. The 1st and 99th percentiles of the transformed χ^2 distribution with 7 degrees of freedom are -3.66 and 1.82 respectively. Hence the Y values should fall between -4 and $+2$ if there are no systematic biases, and the standard deviations

Table 5. *Chi-square (χ^2) agreement statistics for individual parameter estimates across experiments—stem/leaf displays of $Y=(7/2)^{1/2} \log_e (\chi^2/7)$*

	Recreated	Extinction	Robust/Resistant
Heavy atoms: Position parameters			
—3			7
—2		87	
—1		7	3
—0	996110	95	8765
0	011133466789	1222233346	001222477
1	033455779	00012246778	003456677
2	5	116	0115
3	0	2	
4			0
Median	0.6 (1.17)	0.6 (1.17)	0.7 (1.21)

Diagonal thermal vibration parameters

—1	5		
—0	7400		
0	6777		9
1	0245667789	179	04679
2	0233566	0056779	67779
3	0246	002222233467899	0124568899
4		00124	01346668
5			5
Median	1.65 (1.54)	3.2 (2.35)	3.55 (2.58)

Off-diagonal thermal vibration parameters

	Recreated	Extinction	Robust/Resistant
—2	900		
—1	98532	8521100	53210
—0	99996653333220	9822	9855553
0	0011677	01223377789	4455789
1	5	0133778	000255
2		0	22224
Median	—0.4 (0.90)	0.25 (1.07)	0.5 (1.14)

Table 5—(Contd.)

Hydrogen atoms: position parameters

—1		8	1
—0	40		40
0	79	112468	2
1	599	1355578	04778
2	03678	368	0124678
3	1	3	
4	17		
5	5		
6	67		
7			
8			0
9			0

Median 2.45 (1.92) 1.4 (1.45) 1.9 (1.66)

Thermal vibration parameters

Median 0.40 (1.11) 0.88 (1.27) 1.02 (1.31)

of the individual parameter estimates are correct. A shift toward large positive values indicates significant disagreement among parameter estimates.

A stem and leaf display is basically a histogram with an additional unit of precision carried by the plotting symbol. Using the recreated heavy atom position parameter display for illustration, the three smallest Y values are -0.9 , -0.9 and -0.6 , while the two largest values are 2.5 and 3.0 . The 113 model parameters are grouped by type in Table 5. For easy comparison the Y units scale is kept constant within a parameter type. Thus, for example, the units range for heavy atom position parameters is -3 to $+4$.

For the heavy-atom position parameters there is little difference in the distribution of agreement statistics for the three types of refinements, with a slight shift toward positive values in all cases. To avoid undue influence from outliers we use the median to summarize this shift. The number in parenthesis next to the stem-and-leaf median value is the corres-

ponding value of $(\chi^2/K)^{1/2}$, which is a typical ratio of parameter estimate error, as measured by the spread among experiments, to refinement standard deviation. Thus, the standard deviation of heavy-atom position parameters suggest systematic differences of about 20% of the standard deviation for all refinement methods. For heavy-atom diagonal thermal vibration parameters introduction of extinction in the model has a strong influence on the spread of parameters among experiments. Our hope was that including extinction in the model would improve parameter agreement, in the sense that it would only affect the parameter estimates in experiments with large extinction effects. Clearly, extinction has increased the spread of parameter estimates. There must be major model inadequacies, involving correlation between extinction and thermal parameters, that are inconsistent across experiments. Clearly the standard deviations computed in the extinction and robust/resistant refinements tell us little about the accuracy of diagonal thermal vibration parameters.

For off-diagonal thermal vibration parameters the refinement standard deviations are in good agreement with the total spread of the parameters across experiments. There are no strong suggestions of systematic effects in these parameters.

For hydrogen atom position parameters inclusion of extinction eliminates a few wild recreated refinement parameter estimates. The robust/resistant refinement introduces several wild estimates. With a factor of about 1.5 for extinction and robust/resistant refinements, there are still systematic effects for hydrogen atom position estimates which are not related to extinction. Stem-and-leaf displays are not too informative for the six hydrogen atom thermal vibration parameters. The estimates are essentially meaningless because of the large standard deviation estimates, which hide all but the most extreme systematic effects. For six of the eight good experiments robust/resistant standard deviation estimates are

approximately 10% smaller than comparable extinction ones. For experiments 5 and 14 they are 4% larger. This slight systematic effect accounts for the small increase in the typical ratio values for the robust/resistant agreement statistics over that for extinction. Thus, on an absolute basis there are no major differences between extinction and robust/resistant parameter agreement.

Our overall conclusion from Table 5 is similar to those of HA and Mackenzie (1974). There are strong systematic effects in at least some of the experiments which invalidate the refinement standard deviation estimates as measures of the overall precision of parameter estimates. The situation is worst for heavy-atom diagonal thermal vibration parameters, where, typically the refinement standard deviation estimates are low by a factor of 2.5. For hydrogen position parameters the standard deviations are low by a factor of 1.5, and for hydrogen thermal vibration parameters the standard deviations are low by a factor of 1.3. Further, the introduction of extinction in the model causes diagonal thermal vibration parameters to be more variable. Lastly, there are no significant differences between extinction and robust/resistant parameter agreement.

In an attempt to get a better understanding of the nature of the systematic effects we examined the actual changes in the 113 structural parameters (excluding extinction and scale, which are sample dependent) for the three refinement procedures and the eight 'good' experiments. The results varied widely among experiments. However, except for a possible tendency for carbon and oxygen to move by small amounts in opposite directions when the extinction parameter was introduced, the only consistent trend was for the values of the heavy atom diagonal thermal parameters to increase in the extinction refinement. This is a result of the well-known tendency for extinction to depress the intensities of low angle reflections more than high angle ones. Experiment 5

stands out because it appears to show very little extinction, and the shifts of all parameters were generally less than one standard deviation. Experiment 4, on the other hand, had large shifts in both refinement stages. Experiments 9 and 14 also had relatively small changes, but experiment 14 differed from all the others in that the increase in the diagonal thermal parameters of the heavy atoms occurred in the robust/resistant refinement. The results agree with the conclusion of Mackenzie (1974) that extinction was an important source of variability among experiments. It is clear, however, that there are other, still unidentified, systematic effects that influence the different experiments in different ways.

Conclusions

The data from the Single Crystal Intensity Project clearly contain systematic effects that differ among the experiments. We note that when these systematic effects are minimal, or at least are uniform in the full data set, the robust/resistant refinement agrees well with the classical, fully weighted, least-squares refinement. Even when there is good agreement, however, the variation of parameter estimates across experiments is greater than would be expected from the standard deviations calculated by the refinement program. This indicates that the calculated standard deviations are likely to represent an overly optimistic assessment of accuracy unless the data set has been determined, by some independent criterion, to be free of systematic effects.

Several general conclusions may be drawn from the results of this study. *First*, the robust/resistant algorithm converges in some cases where the classical, Newtonian procedure does not. This is a problem in numerical analysis rather than in statistics. While there are other numerical algorithms that are more stable than Newton's method (see, for example,

Broyden, 1972), the robust/resistant algorithm is easy to implement, and has other side benefits. *Second*, with a good data set the robust/resistant procedure gives parameter estimates close to those given by the classical procedure, and there is a strong indication of important systematic effects if the two procedures disagree. Of course, with real data one cannot determine which analysis is closer to the 'correct' structure, because it is not known what the correct structure is. Some comparison can be done with a synthetic data set constructed from a known model (Prince & Nicholson, 1982). *Third*, Huber's (1973) procedure for estimating the standard deviation of a unit weight observation and the standard deviations of parameter estimates gives results that are close to the results given by the classical procedure if the model is adequate.

These conclusions are similar to those obtained by statisticians in using robust/resistant approaches in the fitting of other types of data (Andrews, 1974; Cook, 1977; Mallows, 1979). The robust/resistant analysis reproduces, by a more or less objective procedure, the results of a classical analysis done with a good deal of hand screening of individual observations. It identifies those particular data points that are most inconsistent with the model, which can be used as a starting point for examining both model and data to determine the source of the discrepancy.

References

- ABRAHAM, S. C., HAMILTON, W. C. & MATHIESON, A. McL. (1970). *Acta Cryst.* **A26**, 1-18.
- ANDREWS, D. F. (1974). *Technometrics*, **16**, 523-531.
- BEATON, A. E. & TUKEY, J. W. (1974). *Technometrics*, **16**, 147-185.
- BROYDEN, C. G., (1972). *Numerical Methods for Unconstrained Optimization*, edited by W. Murray, pp. 87-106. London and New York: Academic Press.
- BUSING, W. R., MARTIN, K. O., & LEVY, H. A. (1962). USAEC Oak Ridge National Laboratory Report TM-305.

- COOK, R. D. (1977). *Technometrics*, 19, 15–18.
- COX, G. S., SABINE, T. M. & TAYLOR, G. H. (1966).
- CROMER, D. T., & MANN, J. B. (1968). *Acta Cryst.* A24, 321–324.
- FINGER, L. W. & PRINCE, E. (1975). *NBS Tech. Note* 854.
- HAMILTON, W. C. & ABRAHAMS, S. C. (1970). *Acta Cryst.* A26, 18–24.
- HUBER, P. J. (1973). *Ann. Stat.* 1, 799–821.
- International Tables for X-Ray Crystallography* (1962). Vol. III, Birmingham: Kynoch Press.
- MACKENZIE, J. K. (1974). *Acta Cryst.* A30, 607–616.
- MALLOWS, C. L. (1973). 'On Some Topics in Robustness', paper delivered at the Eastern Regional IMS Meeting, University of Rochester.
- MALLOWS, C. L. (1979). *Am. Stat.* 33, 179–185.
- NICHOLSON, W. L. (1974). Critical Evaluation of Chemical and Physical Structural Information, National Academy of Sciences, Washington, DC, 45–47.
- OKAYA, Y., STEMPLE, N. R. & KAY, M. I. (1966). *Acta Cryst.* 21, 237.
- PRINCE, E. & NICHOLSON, W. L. (1982) to be published.
- STERN, F. & BEEVERS, C.A. (1950). *Acta Cryst.* 3, 341.
- TUKEY, J. W. (1974). Critical Evaluation of Chemical and Physical Structural Information, National Academy of Sciences, Washington, DC, 3–14.
- WELSCH, R. E. & KUH, E. (1977). Massachusetts Institute of Technology and NBER Computer Research Center, WP-923–77.
- ZACHARIASEN, W. H. (1968). *Acta Cryst.* A24, 212–216.

Calculation of the Electron-Density Distribution with an Account of Statistical Errors in Structure Amplitudes and Series Termination

BY A. A. SHEVYREV AND V. I. SIMONOV

Institute of Crystallography, Academy of Sciences of the USSR, Moscow, USSR.

Abstract

On calculating the electron-density distribution in crystals it is desirable to eliminate statistical errors in the observed moduli of structure amplitudes and to smooth out the effect of the Fourier-series termination. Of special importance is the location of light atoms in the presence of heavy ones in the structure as well as the calculation of the difference-density distribution. Proceeding from the mathematical methods of stable Fourier-series summation used when the Fourier-series coefficients are not free from statistical errors (Tikhonov & Arsenin, 1974), the following expression has been derived

$$\rho(\mathbf{r}) = \frac{1}{V} \sum_{\mathbf{H}} w_{\mathbf{H}} F_{\mathbf{H}} \exp [-2\pi i \mathbf{H}\mathbf{r}]$$

where

$$w_{\mathbf{H}} = |F_{\mathbf{H}}|^2 / (|F_{\mathbf{H}}|^2 + \sigma_{|F_{\mathbf{H}}|}^2), \text{ if } |F_{\mathbf{H}}| > \beta \sigma_{|F_{\mathbf{H}}|} \\ = 0, \text{ if } |F_{\mathbf{H}}| \leq \beta \sigma_{|F_{\mathbf{H}}|}.$$

The parameter β depends on the error distribution in $|F_{\mathbf{H}}|_{\text{obs}}$. If the errors follow the Gaussian distribution, $\beta = 2$ is recommended.

The use of special $\sigma_{\mathbf{H}}$ factors was suggested to smooth out the Fourier-series termination waves (Lantsosh, 1961)

$$\rho(\mathbf{r}) = \frac{1}{V} \sum_{\mathbf{H}} \sigma_{\mathbf{H}} F_{\mathbf{H}} \exp [-2\pi i \mathbf{Hr}]$$

In the case of three-dimensional series, the $\sigma_{\mathbf{H}}$ factors have the form

$$\sigma_{hkl} = \frac{\sin[\pi h/(H+1)] \sin[\pi k/(K+1)] \sin[\pi l/(L+1)]}{\pi^3 hkl / [(H+1)(K+1)(L+1)]},$$

where hkl are the usual indices of the corresponding structure amplitudes, whereas the values of HKL depend on the limits of the observed set of F_{hkl} and are determined for each structure amplitude in the following way

$$H = \max h \text{ for the given } k, l \text{ and } h > 0$$

$$H = \max |h| \text{ for the given } k, l \text{ and } h < 0.$$

The values of K and L are determined by the same method. Thus, the $\sigma_{\mathbf{H}}$ factors are peculiar to each Fourier coefficient and depend on its indices and on the used set of structure amplitudes. The full paper gives examples of practical application of the above-mentioned methods, and is published elsewhere (Shevyrev & Simonov, 1981).

References

- LANTSOSH, K. (1961). *Prakticheskie Metody Prikladnogo Analiza* p. 230 Moskva: Fizmatgiz.
 SHEVYREV, A. A. & SIMONOV, V. I. (1981). *Kristallografija*, **26**, 36-41.
 TIKHONOV, A. N. & ARSENIN, V. JA. (1974). *Metody Reshenija Nekorrektnykh Zadach*. Moskva: Nauka.

On Data Reduction and Error Analysis for Single-Crystal Diffraction Intensities

BY ROBERT H. BLESSING AND GEORGE T. DETITTA

*Medical Foundation of Buffalo, Inc., 73 High Street,
Buffalo, NY 14203, USA*

Abstract

Crystallographic studies aimed at detailed mapping of the electron density in molecules and crystals require unusually careful efforts to eliminate systematic experimental errors and to recognize and minimize random errors. Several methods for estimating Bragg peak limits in step-scanned reflection profiles have been developed: minimization of $\sigma(I)/I$ (Lehmann & Larsen, 1974); location of the changes from decreasing peak intensity to 'probably constant' background intensity (Grant & Gabe, 1978); and minimization of an autoconvolution of the intensity profile (Rigoult 1979). These methods become less reliable as peak-to-background values diminish, but, given limits for a suitable sample of the prominent peaks in a data set, anisotropic reflection width parameters can be found by least-squares fit and used to calculate peak limits for even the weakest reflections. To observed base widths W_1 and W_2 below and above the centroids of the 'good' peaks, we fit coefficients q_{ijk} and T_i according to

$$W_i = Q_i + T_i \tan \theta, \quad i = 1, 2,$$

$$Q_i = \sum_{j=1}^3 \sum_{k=1}^3 z_j z_k q_{ijk} = z^T q z.$$

The quantities z_j are components along crystal-fixed Cartesian axes of a unit vector normal to the incident and diffracted beams. For diffractometer axes defined

as in the *International Tables for X-ray Crystallography* (Vol. IV, pp. 276–278), the z_j are given by

$$(z_1, z_2, z_3) = (\sin \phi \sin \chi, \cos \phi \sin \chi, \cos \chi).$$

Our estimates of $\sigma^2(I)$ include contributions from (1) the Poisson variance of the stepwise count rates, corrected for coincidence losses; (2) the variance of the measured dead time of the counting chain, and the variance of the correction factor for the beam attenuator, if used; (3) the variances and covariance of the parameters of a straight line fitted to the background; (4) the variances and covariances of the parameters of polynomial scaling functions of X-ray exposure time fitted to the periodically measured reference intensities; (5) the mean-square deviation from the mean of the scaling factors derived from these functions; and (6) the instrumental variance $p^2 I^2$ (McCandlish, Stout & Andrews, 1975).

Since the Ottawa meeting we have discovered a more correct formulation for the widths of the Bragg peaks as an anisotropic property of the specimen crystal (*cf.* Nelmes, 1980):

$$W_i^2 = Q_i + T_i [\tan \theta(\bar{\alpha})]^2, \quad i = 1, 2$$

where W_1 is the base width of the half peak below $\theta(\alpha_1)$ and W_2 is the base width of the half peak above $\theta(\alpha_2)$, and the Q_i are as defined above.

The research was supported by NIH Grant No. AM-19856. The full paper will be submitted to the *Journal of Applied Crystallography*.

References

- GRANT, D. F. & GABE, E. J. (1978). *J. Appl. Cryst.* **11**, 114–120.
 LEHMAN, M. S. & LARSEN, F. K. (1974). *Acta Cryst.* **A30**, 580–584.
 MCCANDLISH, L. E., STOUT, G. H. & ANDREWS, L. C. (1975). *Acta Cryst.* **A31**, 245–249.
 NELMES, R. J. (1980). *Acta Cryst.* **A36**, 641–652.
 RIGOULT, J. (1979). *J. Appl. Cryst.* **12**, 116–118.

On the Problem of Secondary 'Least-Squares' Minima

BY R. ROTHBAUER

IBM Thomas J. Watson Research Center, Yorktown Heights, New York 10598, USA

Abstract

The method of least squares defines the solution of a system of overdetermined physical equations by an extremal principle. As an immediate consequence of this arises the problem of secondary minima in the related numerical solution algorithms. We therefore develop an alternative to the extremal principle which is not affected by secondary minima. It will be shown that the solutions defined by the new principle differ only negligibly from those derived by the method of least squares if the accuracy of the experiment and of the underlying physical theory are in accordance and that the same statement holds as well for the variance-covariance matrix and for the 'error of an observation with unit weight'. It will also be shown that the results obtained from both principles coincide exactly if the system of physical equations is linear.

Outline

The theories of physics describe nature by irrational numbers representing lengths, times, charges, forces, *etc.* Considering any special object, some of these quantities are in general simply defined by procedures of measurement, while another part is related to the observations by more or less complex theories. The physical quantities of any object under consideration may therefore roughly be separated into observational parameters and model parameters, some of the latter,

$x_{01}, x_{02}, \dots, x_{0p}$, say, being in general determined by some of the first, $y_{01}, y_{02}, \dots, y_{0n}$, say, and a system of equations

$$0 = g_j(y_{01}, y_{02}, \dots, y_{0n}, x_{01}, x_{02}, \dots, x_{0p}), j=1, 2, \dots, l \quad (1)$$


arising from the laws

$$0 = g_j(y_1, y_2, \dots, y_n, x_1, x_2, \dots, x_p), j=1, 2, \dots, l \quad (2)$$

of a theory applied to the object, which states, that in an n -dimensional space B of observations $[y_1, y_2, \dots, y_n]$, only a m -dimensional subspace N is to be observed. If the equations (2) are independent

$$m = n + p - l. \quad (3)$$

Although the theory treats the observables as sharp irrational numbers, they are in practice defined by a prescription of measurement and are thus necessarily of a certain unsharpness, which is commonly assumed to be caused by experimental errors, distributed according to a 'law of errors' if the experiment is repeated. The error concept is used to explain the fact that any result $[y_{01}, y_{02}, \dots, y_{0n}]$ of an experimental investigation will almost never belong to the subspace N of the space B of observations as required by the theory, but will be situated somewhere in its neighbourhood, which means, that the system of equations (1) has generally no exact solution.

What properly should be understood as a solution in the sense of physics is developed in the form of an extremal principle by the 'method of least squares', which describes how the theories of physics are to be applied, and hence is of enormous importance. 

The 'method of least squares' takes the theory, eq. (2), for granted in its full sharpness, assumes that the observations $[y_{01}, y_{02}, \dots, y_{0n}]$, are affected by errors, which are supposed to be distributed with finite standard deviations $[\sigma_1, \sigma_2, \dots, \sigma_n]$ and concludes

that the best approximate solution $[x_{01}, x_{02}, \dots, x_{0p}]$ of (1) is defined by the point $[y'_{01}, y'_{02}, \dots, y'_{0n}]$ on the m -dimensional subspace of the theory—where

$$0 = g_j(y'_{01}, y'_{02}, \dots, y'_{0n}, x_{01}, x_{02}, \dots, x_{0p}),$$

$$j = 1, 2, \dots, 1$$

is mathematically solvable—for which

$$\sum_{j=1}^n (y_{0j} - y'_{0j})^2 / \sigma_j^2 = \text{minimum.} \quad (4)$$

Detailed discussions may be found in the textbooks.

In many cases of practical interest the subspace N of the theory, described by equation (2), and the observation $[y_{01}, y_{02}, \dots, y_{0n}]$, is of such a kind that, besides the main minimum of (4), there exist a great number of secondary local minima without physical significance, where the search algorithms for the unknown model parameters of the 'method of least squares', starting from some guess to be made, may end in practice without a correct solution, if no approximation close enough to the main minimum is known in advance.

The problem of secondary minima severely restricts the possibilities of applying the 'method of least squares'. It is—as we will see later—closely related to an inconsistency in its propositions, which are unsharp observations and theories formulated by strictly valid equations.

The assumption that the observational quantities are subject to probabilistically distributed errors implies that the probability of an observation, which verifies a theory of type (2) with $m < n$, is zero. In other words, the probability of a 'least squares' solution is infinitely small (Gauss, 1839).

If the theories of physics are established from observations by the induction principle, they may only be described by equations if one allows for some unsharpness.

In the forthcoming paper we will, because of this, drop the assumption of laws described by strictly valid equations (2). On this basis we will develop two alternative definitions for the solution of (1), which are not affected by the problem of secondary minima. We will show that the deviations of the solutions given by these definitions from those given by the extremal principle of the 'method of least squares' are negligible if the problem (1) is physically relevant. In particular it will appear that all three definitions coincide if the system of equations (1) is linear.

The full paper is available as *IBM Research Report* RC 9045 (39595) Mathematics, and will be submitted for publication elsewhere.

Reference

GAUSS, C. F. (1839). Letter to Bessel dated 26 February.

Wiener Methods for Electron Density

BY D. M. COLLINS AND M. C. MAHAR

*Department of Chemistry, Texas A & M University,
College Station, Texas 77843, USA*

Abstract

The Wiener formalism is widely used in applications conveniently categorized as smoothing, interpolation, or extrapolation of stationary series. The present application is of the last-mentioned type and consists in extrapolation of a set of structure factors (phases and magnitudes) beyond the experimental (2θ) limit of data to increase resolution in the corresponding density function. The application has in view cases for which data are severely curtailed in angular range, but not necessarily in number. Biological macromolecular structure problems, though beyond reach at present, fit the application exactly and, in fact, were the target from the beginning.

Suppose a set of structure factors, spherically complete for some range of $|\mathbf{h}|$ including $|\mathbf{h}| = 0$. Now an estimation of electron density at higher resolution than nominally provided by the original structure factors is found by solving the matrix equation

$$\mathbf{FC} = \beta,$$

for \mathbf{C} and computing

$$\rho(\mathbf{r}) \simeq \kappa / |\Sigma C(\mathbf{h}) \exp\{-2\pi i \mathbf{h} \cdot \mathbf{r}\}|^2,$$

where κ is a collection of constants, and the summation is over a half-lattice and its origin at which $C = 1.0$. The determinant $|\mathbf{F}|$ is of the general Karle-Hauptman type but with entries restricted by the half-
C. S.—18

lattice condition. The lead element of β is positive, the others are zero.

1. Introduction

Norbert Wiener's (1949) *Extrapolation, Interpolation, and Smoothing of Stationary Time Series* provides the basis for the material presented. It should be noted at the outset that other parallel work has led to common attribution of prediction theory, which emphasizes the aspect of extrapolation, to both Kolmogorov and Wiener as originators (*Encyclopedic Dictionary of Mathematics*, 1977). In recognition of this, we shall have occasion to refer to Wiener-Kolmogorov (wk, hereafter) prediction theory but it is the work of Wiener which lies behind much of what follows.

It was Wiener's (1949) purpose to bring together the theory and practice of two fields of diverse tradition, communication engineering and time series in statistics. In the latter field algebraic relationships, especially those involving correlations, are used to find desired results which are best in some average sense. Communication engineering has the special concern of discovering the stability of oscillatory systems and whether their oscillations die out or grow without bound. The present crystallographic application employs Wiener's synthesis of the different techniques but in simplified forms worked out primarily by geophysicists (*cf.* Robinson & Treitel, 1980).

Wiener's (1949) synthesis of techniques and the various forms which result are based upon the (one-dimensional) Fourier-transform pair. The appropriate relationships are

$$\rho(x) = \int_{-\infty}^{+\infty} F(u) \exp \{-2\pi i u x\} du, \quad (1)$$

which defines $F(u)$, and $F(u)$ is given by

$$F(u) = \int_{-\infty}^{+\infty} \rho(x) \exp \{2\pi i u x\} dx. \quad (2)$$

In crystallographic application the continuous but periodic electron density ρ requires a discretely sampled structure factor F and, in fact, the current fast Fourier-transform computer algorithms require both ρ and F to be discretely sampled on regular grids (Gentleman & Sande, 1966) for numerical calculations. Relaxation of the customary presumption that Fourier analysis is to be restricted to the axis of real values of x constitutes an apparent complication as ρ thus becomes a function in the complex plane. But it is advantageous to move off the axis of reals into the complex plane because ρ then can be subjected to the full power of analytic-function theory. This is the approach of Wiener and in his book-length analysis he presents the detail necessary for mathematical rigour, but with emphasis upon functions which are continuous rather than discretely sampled as in crystallographic applications.

The present book deals with statistics in a very general sense and it is worth emphasizing that structure factors (both amplitude and phase) are used quite freely in this part as numerical facts, as statistics. The emphasis is not on algebraic relationships between structure factor and electron density, but on the density itself, or at least one of its roots, as a function of definite character in agreement with the available statistical information, the structure factors. The criterion of agreement is the statistical measure of mean-squared-error whose functionally constrained minimization is often the means to a desired result.

Gassmann (1977) discussed Wiener filtering as structure determination by the filtering of noisy images. In that paper a filter was given for which the denominator is a Patterson coefficient. From an algebraic point of view it is necessary only to make certain that no individual Patterson coefficient is zero in order that the filter be valid. In functional analysis the proposal of such a filter constitutes a very daring assertion indeed. It is the relevant analysis to which Wiener (1949) had addressed himself and from

which the form and power of his methods are derived. Functions free of zeros and analytic along an axis of reals and in a contiguous region of the complex plane have the central role in Wiener's work. Let it be observed that in each of the following applications, including the connections with information theory, a result always turns upon implicit or explicit discovery of functions in n -dimensional generalization which are similar to those used by Wiener.

2. Wiener's problem and its formal solution

Wiener's (1949) simplest and basic problem was the extrapolation or prediction of time series. A prediction, of course, can never be a perfect continuation of a series for such a state of affairs would preclude the possibility of new information becoming available. Nevertheless, a series is subject to statistical prediction and the problem may be posed as the prediction or estimation of data not yet measured. The series to be treated are assumed real (for convenience) and stationary which in regard to the autocorrelation

$$\phi_{\tau} = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{t=-N}^N x_{t+\tau} x_t, \quad (3)$$

may be interpreted to mean not only that ϕ exists, as it must for a physical process, but also that

$$\phi_{\tau} = \lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{t=-N}^N x_{t-Q+\tau} x_{t-Q}, \quad (4)$$

where Q is arbitrary and ϕ_{τ} is entirely unaffected by its specification. It will, of course, be most convenient to take $Q = 0$ but the formulation makes it clear that an autocorrelation is independent of the time origin for a stationary series.

Suppose a series known for all past time and trun-

cated at t . The prediction problem is solved upon successful estimation of future series members as

$$\hat{x}_{t+a} = a_0 x_t + a_1 x_{t-1} + a_2 x_{t-2} + \dots, \quad (5)$$

for which there is a minimum in the formal expression

$$\lim_{N \rightarrow \infty} \frac{1}{2N+1} \sum_{t=-N}^N |x_{t+a} - \hat{x}_{t+a}|^2. \quad (6)$$

It is clear that the coefficients a of a linear prediction operator are generally dependent upon the prediction span a . For any a the prediction operator is obtained by solving the equations resulting from minimization of the time expectation

$$I = E \left\{ \left[x_{t+a} - \sum_{s=0}^{\infty} x_{t-s} a_s(a) \right]^2 \right\}, \quad (7)$$

with respect to $a_s(a)$. The equations to be solved are

$$\sum_{s=0}^{\infty} a_s(a) \phi_{\tau-s} = \phi_{\tau+a}; \quad \tau = 0, 1, 2, \dots \quad (8)$$

and the resulting WK linear predictor may be used in (5) to obtain estimates of data not yet measured.

An obvious and important feature of (8) is that the WK linear predictor depends only upon the autocorrelation of a process. The application of a prediction operator is independent of origin definition in time and, in fact, is even independent of the choice of time series itself so long as it is chosen from the infinity of series with autocorrelation coefficients equal to those used in (8). Although the summation index in (8) takes on only non-negative values, it could take on all values because by construction a is a causal or one-sided function which vanishes for negative values of s .

If one had a reasonably large set of autocorrelation coefficients without lacunae, (8) would lead at once to an approximation of (5) and a formally satisfactory solution to the simple prediction problem. Alternatively, Fourier coefficients for a might be sought for application in the dual space through transformation of (8). Either approach is problematic because a physical process and its autocorrelation are not measured simultaneously. This practical principle of complementarity dictates the inaccessibility to physical measurement of either according to the measurability of the other. The common side of the principle is that for which a process is measured and its autocorrelation must be approximated by some method of spectral estimation (Oppenheim & Schaefer, 1975). This is a very substantial problem and wk prediction often founders upon the difficulty of autocorrelation determination.

Crystallographic application (Collins, 1978) involving extrapolation in reciprocal space, hence resolution enhancement in direct space, is a problem of the other sort. Electron-density must be positive definite and may therefore be written as

$$\rho = |g|^2 > 0 \quad (9)$$

with complete generality. Fourier transformation of (9) shows the autocorrelation of G , the transform of g , is F and the resolution-enhancement problem is seen to begin with the knowledge of an autocorrelation and the need for its deconvolution. But the difficulties of wk extrapolation are not removed by this change of complexion. The deconvolution of F is as much a problem as the convolution of x and the extrapolation of G as uncertain as that of x apart from other information or constraints.

The needed constraints are provided in Wiener's (1949) development of the wk linear predictor through functions analytic and free of zeros and poles in a half-plane and on its (finite) boundaries. For resolu-

tion enhancement the corresponding computations would involve explicit extrapolation of G and subsequent convolution to yield estimates of high-resolution structure factors. As an explicit computation this is entirely impractical but it is achieved by implication in § 3.

3. Density and filters in one dimension

A. Principles. Wiener's (1949) analyses were concerned with continuous functions almost exclusively. For crystallographic applications most Fourier transformations involve discretely sampled functions, both density and its transform. The corresponding implied periodicity in both spaces invites Fourier analysis upon the unit circle or its n -dimensional generalization, the unit polycylinder. An immediate practical result is the z -transform obtained for time series by replacing $\exp(-2\pi if)$ with z . Then a Fourier transform and the related z -transform are

$$X(f) = \frac{1}{L} \sum_{t=-\infty}^{\infty} x_t \exp \{-2\pi ift\}, \quad (10)$$

$$X(z) = \frac{1}{L} \sum_{t=-\infty}^{\infty} x_t z^t. \quad (11)$$

While (11) has been written with $|z| = 1$, it is clear that restriction of X to be on the unit circle is necessary only to preserve its nature as a Fourier transform. $X(z)$ is a Laurent series at every (finite) point in the complex plane and upon the unit circle is the periodic Fourier transform of x . Let it be observed that the change of variable maps the half-plane and its boundaries over which Wiener's functions would be free of zeros and poles to the unit circle and the interior domain which it bounds.

Error of extrapolation [*cf.* equation (5)] is readily composed for $\alpha = 0$ as

$$\epsilon_t = x_t - \hat{x}_t = x_t - \sum_{s=-\infty}^{t-1} x_s a_{t-s}, \quad (12)$$

$$1 - a = (1, -a_0, -a_1, \dots) = \gamma. \quad (13)$$

It is assumed that only x is known and that through its Fourier transform it satisfies the Paley-Wiener condition (Collins, 1978; Robinson, 1967) which requires $0 < |X(f)| < \infty$. Determination of γ is to be driven by constraints upon it and upon the error process ϵ . An initial motivation for determination of the unit-extrapolation-error (or prediction-error) filter γ is its application by (13) and (5) to unit extrapolation of x .

Another equally important use for γ follows from the constraint placed upon ϵ . What seems the most cogent constraint upon ϵ in (12) is that the error series should be uncorrelated (Collins, 1978; Robinson & Treitel, 1980). If it were not so, then the filter γ would not have extracted the stationary statistical features or predictability from x . The constraint is imposed upon the z -transform of

$$\epsilon_t = \sum_{s=-\infty}^{\infty} x_s \gamma_{t-s}, \quad (14)$$

$$E(z) = X(z) \Gamma(z), \quad (15)$$

for which

$$|E(f)|^2 = |X(f)|^2 |\Gamma(f)|^2 = \sigma^2/L; \quad (16)$$

σ^2 is a constant and L is the period of f . This equation expresses the requirement that ϵ be an uncorrelated error series for only if this is so is its Fourier transform of constant magnitude as required by (16). Evidently $1/|\Gamma(f)|^2$ gives the spectral characteristics of x and apart from a constant factor completely represents its correlations and their structure.

Applications require the practicality of finite filters for which (14) may be rewritten as

$$\epsilon_t = \sum_{s=t-n}^t x_s \gamma_{t-s}. \quad (17)$$

The change in upper limit is only a formality as γ was designed to be causal or one-sided; the change in lower limit reflects an arbitrary limit of $n + 1$ upon filter length and the requirement that higher filter elements be zero. Because there is no limit to the length of x in time past, (16) may be written as

$$\sigma_n^2/L = |X(f)|^2 \left| \sum_{s=0}^n \gamma_s \exp \{-2\pi i s f\} \right|^2 \quad (18)$$

in conformity with (17). Presumably ϵ does not vanish and since for a physical process of the sort under consideration $0 < |X(f)| < \infty$, then

$$\Gamma(z) = \sum_{s=0}^n \gamma_s z^s \neq 0, \quad |z| = 1, \quad (19)$$

and equation (18) may be rearranged to

$$|X(f)|^2 = \frac{\sigma_n^2/L}{\left| \sum_{s=0}^n \gamma_s \exp \{-2\pi i s f\} \right|^2}, \quad (20)$$

the form in which the spectral density of x depends only upon the causal extrapolation-error filter, apart from a constant factor.

The extrapolation-error filter is determined by use of equations (8) and (13). A unit prediction span and a filter length of $n + 1$ lead to

$$\begin{bmatrix} \phi_0 & \phi_{-1} & \dots & \phi_{-n} \\ \phi_1 & \phi_0 & & \\ \vdots & & & \\ \phi_n & \phi_{n-1} & \dots & \phi_0 \end{bmatrix} \begin{bmatrix} \gamma_0 = 1 \\ \gamma_1 \\ \vdots \\ \gamma_n \end{bmatrix} = \begin{bmatrix} \beta \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (21)$$

as is easily verified. Because x may be combined with its Hermitian transpose to give $\phi = L^{-1} x^\dagger x$, equation (14) may be used to equate β and the variance of ϵ by

$$\sigma^2 = \frac{1}{L} \epsilon^\dagger \epsilon = \frac{1}{L} \gamma^\dagger x^\dagger x \gamma = \gamma^\dagger \phi \gamma = \beta \geq 0. \quad (22)$$

Equation (21) is solvable in the ordinary sense for any order n if the matrix ϕ is positive definite. The equation is discussed at length in the geophysical literature in which it is shown to be very generally solvable, even when ill-conditioned, by a highly efficient recursion procedure based upon the Toeplitz structure of ϕ (Claerbout, 1976; Robinson & Treitel, 1980).

The extrapolation-error filter is unique and in two senses. First, within a constant factor $|\Gamma(z)|^2$ is given on the unit circle by $|X(f)|^{-2}$. By function theory the ordinary polynomial (19) has on the unit circle its maximum and minimum moduli in the region bounded by the unit circle (Carathéodory, 1964). Thus $\Gamma(z)$, having no zeros on or within the unit circle, has all its roots at $|z| > 1$. But with $|\Gamma(z)|^2$ given for $|z| = 1$ this corresponds to the unique Fejér factorization in which all and only the roots of $|\Gamma(f)|^2$ outside the unit circle are used in the formation of $\Gamma(z)$ (Collins, 1978). Evidently γ , the extrapolation-error filter, the Fourier transform of unique $\Gamma(f)$, is itself unique. Second, γ is a minimum-delay operator. Consider some $\gamma' \neq \gamma$ for which $|\Gamma'(f)|^2 = |\Gamma(f)|^2$. The minimum-delay property is given by Robinson's theorem (Claerbout, 1976)

$$\sum_{s=0}^m |\gamma_s|^2 \geq \sum_{s=0}^m |\gamma'_s|^2; \quad m = 0, 1, \dots, n. \quad (23)$$

For at least one of the equations given by (23) the equality must fail if $\gamma' \neq \gamma$. This minimum-delay property of unique γ shows that there is no other

sequence of coefficients which with equal compactness represents $|\Gamma(f)|^2$, hence $|X(f)|^2$ and the statistical structure of stationary x .

B. Applications. Crystallographic application to resolution enhancement in one dimension follows from the assertion of equation (9) that $\rho = |g|^2 > 0$. As pointed out earlier, this requires F to be an auto-correlation and there follows immediately an equation analogous to (21).

$$\begin{bmatrix} F_0 & F_1^* & \dots & F_n^* \\ F_1 & F_0 & & \\ \vdots & & & \\ F_n & F_{n-1} & \dots & F_0 \end{bmatrix} \begin{bmatrix} C_0 = 1 \\ C_1 \\ \vdots \\ C_n \end{bmatrix} = \begin{bmatrix} \sigma_n^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \quad (24)$$

$$c_{\rho_x} = \frac{\sigma_n^2/L}{n \left| \sum_{k=0}^n C_k \exp \{ -2\pi i k x \} \right|^2}, \quad (25)$$

a high-resolution electron-density estimator (Collins, 1978); L is the period of both ρ and g .

There is a variety of ways to see that c_{ρ} is a high-resolution estimator of density. The most cogent evidence is given in Robinson's theorem. This theorem requires that C be the most compact representation of ρ in the sense that for any limit of coefficient order, C will represent ρ as well or better than any other set of coefficients, even F .

Perspective on this remarkable fact is provided by practical consideration of the solution of (24). It is known that the Karle-Hauptman (1950) determinants approach singularity in the neighbourhood of order N , the number of atoms in a unit cell (Goedkoop, 1950), and therefore the number of non-trivial elements in C is limited to being not much larger than N . Clearly, if the elements of a set of structure factors are exceedingly more numerous than the atoms in a unit cell, then normal Fourier synthesis will yield an authoritative representation of electron density which cannot be matched by c_{ρ} . If, on the other hand, the

number of known structure factors is less than N , c_ρ is a density estimate at a resolution higher than that provided by Fourier synthesis of structure factors. In connection with information theory this phenomenon is called superresolution (Benjamin, 1980).

These two extremes are not representative of the largest part of crystallographic problems either in the somewhat abstract one-dimensional case of this section or in three dimensions. A desirable but so far unknown analysis for a representative real case would combine quantitatively the effects of additional or imposed information and experimental error to give an estimate of potential resolution enhancement in c_ρ . In the absence of such analysis the nature of these effects is nevertheless clear. The two outstanding facts of principle are that resolution in c_ρ is without limit, and experimental error in any particular F is distributed over all elements of C in some manner consistent with positive-definite density. Both facts, it should be noted, follow directly from the functional constraints required for discovery of unique, analytic $1/g$ free of zeros and poles on and within the unit circle as required by WK prediction theory.

While a formalism for resolution enhancement has been presented, its measure has been achieved only in particular cases through study by simulation. A one-dimensional crystallographic illustration discussed in detail elsewhere (Collins, 1978) here provides an empirical basis for expectation of resolution enhancement by factors > 2 in favourable cases. The artificial structure described in Table 1 was constructed to contain atoms separated by typical interatomic distances. The computations are all based on error-free structure factors calculated for atoms free of thermal motion. Solution of equation (24) at order $n=14$ gave the elements of C tabulated in Table 2. The corresponding density functions, c_ρ and ρ based upon Fourier synthesis of structure factors with indices in the range 0-14, are given in Fig. 1. In the figure c_ρ is plotted above the zero line and increas-

Table 1. *Atomic position parameters and interatomic distances for a hypothetical one-dimensional structure*

Atom	x	Inter-atomic distance, Å	Atom	x	Inter-atomic distance, Å
C ₁	0.01094	1.5	N ₈	0.45312	1.3
O ₂	0.03438		C ₉	0.47344	
C ₃	0.05312	1.2	O ₁₀	0.70938	15.1
N ₄	0.07500	1.4	C ₁₁	0.72969	1.3
C ₅	0.09688	1.4	N ₁₂	0.78906	3.8
O ₆	0.25469	10.1	C ₁₃	0.81250	1.5
C ₇	0.43125	11.3	C ₁	1.01094	12.7
N ₈	0.45312	1.4			—
Axis length = 64.0					

 Table 2. *The extrapolation-error filter derived from data at 4.6 Å resolution,* $C_h = A_h + iB_h$.*

h	A_h	B_h
0	1.000	0.0
1	-0.416	0.146
2	0.068	-0.190
3	0.132	-1.417
4	-0.975	0.449
5	0.242	-1.155
6	-0.598	0.642
7	-0.528	1.536
8	-0.691	-0.447
9	0.624	0.686
10	0.261	0.168
11	0.558	-0.253
12	0.485	0.155
13	-0.246	0.071
14	-0.087	-0.457

 *For this filter, the error-series variance is $\sigma_{14}^2 = 2.94$.

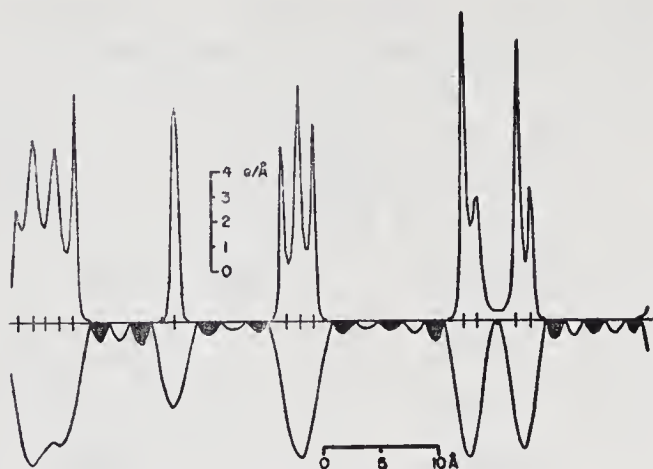


Fig. 1. One-dimensional electron-density functions based on data at 4.6 Å resolution.

ing upward; ρ is plotted below the line and increasing downward. It is clear from the formulation of c_ρ that it is always positive, but ρ , having negative values, is actually plotted as its absolute value with negative areas shaded. The positions of the 13 atoms are marked by vertical hash marks across the zero line.

It is clear that the resolution in c_ρ is greater than in ρ . It is to be noted that the two 1.3 Å separations have been resolved but a 1.2 Å separation has not. The rule of thumb which requires data at a resolution (minimum interplanar spacing) just equal to the distance between atoms to be resolved, in this case implies the effective resolution in c_ρ is in the range 1.2–1.3 Å. The structure-factor data employed are at a resolution of 4.6 Å (minimum interplanar spacing) and this simulation corresponds to qualitative resolution enhancement by a factor somewhat greater than 3. Resolution of the 1.2 Å separation is not achieved until structure factors with indices up to 32 are used (Collins, 1978). The corresponding data resolution is 2.0 Å (minimum interplanar spacing) and the qualitative resolution enhancement is thus reduced to somewhat under 2. In this latter case the order of equation (24) is 32 and is significantly greater than $N=13$.

The corresponding results should not be considered especially reliable in view of the ill-conditioning problem discussed earlier. In consideration of these observations and the known non-linearity of relationship between superresolution and the extent of parent data (Benjamin, 1980), it is reasonable to make the following empirical assertion. Wiener methods may be employed in crystallographic problems to enhance resolution in density functions by a factor >2 under favourable conditions.

The structure factors used in the illustration are error-free. The error-series variance σ_n^2 is nevertheless not zero but gives a measure of the degree to which C does not represent the information contained in F . Because C is unitless, σ_n^2 has the units of F and gives as a number of electrons the global mismatch between ρ , the transform of experimental F , and C_ρ . For $n=1$ there is no structure in C_ρ and its mismatch with ρ is exactly given by the total number of electrons per lattice point. In the real case for which structure factors may be taken as the sum of true values and (small) random numbers, experimental ρ will also be the sum of true values and (small) random numbers. The necessarily zero-mean random process cannot be represented by positive-definite C_ρ and although the errors in F may be distributed unpredictably over C_ρ the value for σ_n^2 will certainly increase as ordinary experimental error is carried into the values for F .

4. Helson, Lowdenslager and n -Dimensions

A. Principles. In the preceding section discussion was limited mainly to one dimension. The transition to two or more dimensions is not straightforward. This section and the n -dimension generalization of Wiener methods depend heavily upon the work of Helson and Lowdenslager (1958). In that paper the authors point to the underlying problem of n -dimensional generalization in the statement that 'analytic function theory

divides into two distinct disciplines in higher dimensions'. Here reference is made to the theory of functions on the unit bicylinder as contrasted to the unit circle. The generalization problem has an obvious practical manifestation in the loss of Toeplitz-form matrices in the two-dimensional form of equation (24). For present purposes the loss of Toeplitz form is the sole problem of generalization from one dimension and the two-dimensional case therefore suffices as a representative for any number of dimensions. For the following discussion many findings and statements have been taken without proof from Helson and Lowdenslager (HL, hereafter) whose presentation provides a full discussion of each point.

The foundational construction which corresponds to the causality or one-sidedness of one dimension is the half-lattice of two (or more) dimensions proposed by HL. On a two-dimensional primitive net whose points are represented by integral coordinates (m,n) , S is a half-lattice if

- a. $(0,0)$ does not exist in S ,
- b. (m,n) exists in S if and only if $(-m,-n)$ does not exist in S unless $m=n=0$.
- c. (m,n) exists in S and (m',n') exists in S imply $(m+m',n+n')$ exists in S .

It may be noted that this definition of a half-lattice is suitable for any number of dimensions including one.

The foundational theorem for n -dimensional generalization of C_p makes use of the half-lattice as a domain throughout which Fourier coefficients may be non-zero. The theorem, a slightly weakened form of HL theorem 2, is:

Let p be summable on the bicylinder and given by the Fourier series

$$p(e^{ix}, e^{iy}) = P_{00} + \sum_S P_{mn} \exp \{-2\pi i(mx + ny)\} \quad (26)$$

where S is any half-lattice. Then

$$\int \ln |p| dx dy \geq \ln |P_{00}|. \quad (27)$$

The point of immediate importance is that $|p| > 0$, provided the half-lattice condition is observed in (26), and $|P_{00}| > 0$. If $|p| > 0$ and is given by (26), then p^{-1} is summable on the unit bicylinder and both it and its absolute value are well behaved and non-zero. A formal two-dimensional electron-density estimator analogous to C_{ρ_x} may be written as

$$C_{\rho_{xy}} = \frac{\sigma_n^2/A}{|C_{00} + \sum_S C_{hk} \exp \{-2\pi i(hx + ky)\}|^2}, \quad (28)$$

where A is the area of the lattice period and, as for the one-dimensional case, $C_{00} = 1$ which here also ensures the finiteness of $C_{\rho_{xy}}$ by HL theorem 2.

Relationships necessary to evaluate (28) are readily developed along the lines used earlier in the one-dimensional case. As before, the key physical assertions are the existence and availability of structure factors and the positivity of electron density expressed by existence of non-zero g such that $\rho = |g|^2 > 0$. HL show that if ρ (real) is non-negative and summable on the bicylinder, then for any definite half-lattice there is a unique H given by

$$H(x, y) = \sum_S C_{hk} \exp \{-2\pi i(hx + ky)\}, \quad (29)$$

such that

$$\rho = (|1 + H|/\kappa)^{-2} \quad (30)$$

where κ is a positive constant. With $c \equiv 1 + H$ it is evidently the case that

$$\kappa |g^{-1}| = |c|, \quad (31)$$

and

$$g c = \kappa \exp \{i\phi\}, \quad (32)$$

where ϕ is undetermined and c has the form given in (26) with $C_{00}=1.0$. The Fourier transform of (32) may be written in matrix notation as

$$\mathbf{G} \mathbf{C} = \epsilon \quad (33)$$

and premultiplication by $A^{-1} \mathbf{G}^\dagger$ gives

$$\frac{1}{A} \mathbf{G}^\dagger \mathbf{G} \mathbf{C} = \mathbf{F} \mathbf{C} = \beta \quad (34)$$

because F is the autocorrelation of G as required by (9). It is interesting that if equation (33) is satisfied and \mathbf{F} is not singular, then

$$\mathbf{C}^\dagger \mathbf{F} \mathbf{C} = \sigma^2, \quad (35)$$

for any suitably chosen elements for β , not all zero, as required by (32) and the consequent random nature of ϵ . Analogy with the one-dimensional case suggests that except for the lead element, all elements of β be taken as zero. It is easily verified that after multiplicative adjustment to make $C_{00}=1$ the lead element of β is σ^2 , the variance of ϵ . Complete determination of \mathbf{C} , subject to specification of a half-lattice, may be obtained as the solution of

$$\begin{bmatrix} F_{00} & F_{01} & \dots & F_{0n} \\ F_{10} & F_{11} & & \\ \vdots & & & \\ F_{n0} & F_{n1} & \dots & F_{nn} \end{bmatrix} \begin{bmatrix} C_0 = 1 \\ C_1 \\ \vdots \\ C_n \end{bmatrix} = \begin{bmatrix} \sigma_n^2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad (36)$$

in which the subscripts \mathbf{h}_i and $\mathbf{h}_i - \mathbf{h}_j$ have been replaced by i and ij respectively.

Practical selection of a half-lattice by its definition is a little awkward. For crystallographic application the problem is greatly simplified by construction of the lattice starting with some reciprocal lattice point \mathbf{h}_0 . All \mathbf{h} , $|\mathbf{h}| < |\mathbf{h}_0|$ may then be added to \mathbf{h}_0 to generate a half-lattice S' . This lattice is somewhat more restricted than HL require, but it satisfies their geometry and in any case may be allowed to increase without bound by allowing \mathbf{h}_0 to become arbitrarily large.

The great advantage of the construction centred on \mathbf{h}_0 arises in the following way. Consider that for some \mathbf{h}_0 the set $(\mathbf{h}'_i = \mathbf{h}_i + \mathbf{h}_0)$ may be listed in order of increasing magnitude of \mathbf{h}_i and that the corresponding elements of \mathbf{C} may be given in the same order, and C_{00} last. Now F_{ij} the elements of \mathbf{F} are independent of \mathbf{h}_0 because $\mathbf{h}'_i - \mathbf{h}'_j = \mathbf{h}_i - \mathbf{h}_j$. Furthermore, in the expression

$$\left| \sum_{T'} C_{\mathbf{h}'} \exp \{ - 2\pi i \mathbf{h}' \cdot \mathbf{x} \} \right|^{-2}, \quad (37)$$

where T' signifies the half lattice S' and its origin, all \mathbf{h}' may be changed by an arbitrary fixed vector, say $-\mathbf{h}_0$, without changing the numerical value of the expression. Because \mathbf{F} is unaffected by \mathbf{h}_0 as is the value of (37), equation (36) may be set up with $\mathbf{h}_0 = (0,0)$ then \mathbf{h}_i , $|\mathbf{h}_i| \geq |\mathbf{h}_{i-1}|$; $i = 1, 2, \dots, n$. It does not appear that omission of any \mathbf{h} need be inimical to evaluation of C_p but it is not clear that there is an advantage to any such omission and it is not considered further.

B. Applications. Crystallographic application involving n -dimensional density estimation requires evaluation of equation (36). Density estimation then follows after evaluating.

$$^c \rho_{\mathbf{x}} = \frac{\sigma_n^2/A}{\left| \sum_T C_{\mathbf{h}} \exp \{ - 2\pi i \mathbf{h} \cdot \mathbf{x} \} \right|^2}. \quad (38)$$

In this equation, which is a modified form of equation (28), T represents a lattice portion including the origin and points well distributed about the origin, $|\mathbf{h}| \leq |\mathbf{t}|/2$. While the maximum of $|\mathbf{h}|$ is $|\mathbf{t}|/2$, the maximum of $|\mathbf{h}_i - \mathbf{h}_j|$ is $|\mathbf{t}|$; $|\mathbf{t}|/2$ is designated 'estimator resolution' when expressed as a minimum interplanar spacing, similarly $|\mathbf{t}|$ is 'information resolution.' It may be noted that in the one-dimensional case the true half-lattice is retained for applications and estimator resolution is the same as information resolution.

A choice of n , the number of elements in T excluding the origin, must be limited to that for which the implied information resolution lies within the range of available structure-factor data. It may be that the limit by data is not absolute, but it represents the actual limit of information employed whether or not n is allowed to increase further. The number of atoms in a unit cell is, as in one dimension, an upper limit for n insofar as (36) tends toward ill-conditioning as n increases beyond N .

Simulation studies were conducted using the two-dimensional centrosymmetric projection of hexamethylbenzene reported by Brockway and Robertson (1939). After minor adjustment of the reported coordinates and determination of an overall isotropic thermal parameter to be $B = 3.0 \text{ \AA}^2$, structure factors corresponding to the observations were calculated. These calculated structure factors were used as error-free data for the simulation studies except as noted. For the structure model $R = 0.21$.

Straightforward evaluation of c_ρ by equations (36) and one equivalent to (38) was carried out for a variety of values for n . Although the atoms, all carbon, were resolved for a smallest value of $n = 52$, the qualitatively similar map for $n = 54$ was subjectively judged to be a significantly better representation of the structure. Fig. 2 gives c_ρ for $n = 54$. Fig. 3 gives ρ on the same scale as computed by

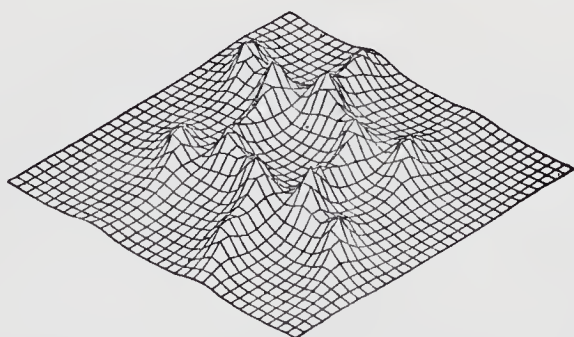


Fig. 2. Hexamethylbenzene projection given by an order 54 electron-density estimator.

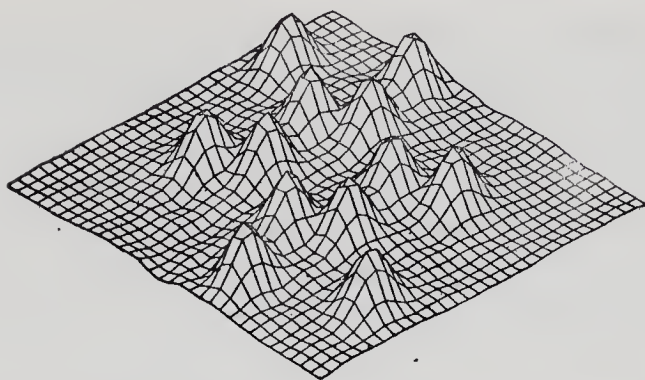


Fig. 3. Hexamethylbenzene projection at high resolution given by Fourier synthesis of error-free structure factors.

normal Fourier synthesis of the 108 unique (perfect) structure factors corresponding to the complete set of reported observations. For $n = 54$, $\sigma_n^2 = 37$ and the estimator resolution is 1.62 \AA or $\sin \theta/\lambda = 0.31 \text{ \AA}^{-1}$; the corresponding information resolution is 0.81 \AA or $\sin \theta/\lambda = 0.62 \text{ \AA}^{-1}$.

The resolution of the original set of data is 0.78 \AA or $\sin \theta/\lambda = 0.64 \text{ \AA}^{-1}$. Clearly, C_ρ as calculated for $n = 54$ does not represent a more efficient use of information than does normal Fourier synthesis. Figs. 2 and 3 show that C_ρ is naturally more spiky than a corresponding Fourier synthesis of the structure factor data upon which C_ρ is based. This parallels the common crystallographic practice of 'sharpening' density functions, whether in electron-density or Patterson maps, and suggests that unmodified structure factors may not be the most suitable for resolution enhancement. Sharpening of any kind is achieved by alteration of structure factors to remove some of or all their global $\sin \theta$ fall-off. Use of altered structure factors in the matrix \mathbf{F} expands the range of its eigenvalues so the smallest become smaller and the determinant of \mathbf{F} tends toward zero. This effect provides the basis for the following test.

Calculations designed to force maximum resolution enhancement in C_ρ used calculated phase and reported experimental structure-factor moduli (over) corrected

for thermal motion by $\exp \{B \sin^2 \theta / \lambda^2\}$ in which $B = 6.0$. For various values of n , a cycle of solutions of equation (36) was devised which would drive $|\mathbf{F}|$ toward singularity. The cycle comprised solution of (36) for a selected value of n , complete eigen-analysis of \mathbf{F} and subsequent reduction of all diagonal elements by the smallest eigenvalue $\times 3/4$ before a new solution of (36). From a series for $n = 8, 12, 16, 20, 24$, C_ρ for $n = 20$ and F_{00} reduced to 69 was judged to be the function which displayed maximum effective resolution enhancement. Each of the six independent atoms appeared in C_ρ at the right location; three were given by single peaks, three by double peaks and there were no other significant peaks in the map. The map is excessively spiky with its highest peak at $77\text{e}/\text{\AA}^2$, its lowest at $20\text{e}/\text{\AA}^2$ and about 80% of the map below $1\text{e}/\text{\AA}^2$. In this case $\sigma_{20}^2 = 16$ and the ratio $\sigma_n^2/F_{00} = 0.23$, a substantial improvement over the earlier case for $n = 54$ in which $\sigma_n^2/F_{00} = 0.51$.

Resolution of the atoms in the projection of hexamethylbenzene undoubtedly could be forced for lower values of n had the data been more exact or the calculations more tightly constrained. For $n = 20$ the estimator resolution is 2.75\AA , the information resolution is 1.37\AA and the smallest interatomic separation resolved is a foreshortened 1.06\AA . Concerning qualitative location of equal atoms, the rule of thumb mentioned in discussion of the one-dimensional case leads to an expectation of potential n -dimensional resolution enhancement of at least $1.37\text{\AA}/1.06\text{\AA} = 1.3$ provided the structure-factor moduli are in error by less than 20%.

5. Connections with information theory

Resolution enhancement as a specific goal in the development and application of Wiener methods for electron density has a clear conceptual parallel with Shannon & Weaver's (1949) information theory.

As presented by Shannon, information theory was concerned with the efficiency of encoding and transmitting information. The crystallographic problem of resolution enhancement is similar but has a different emphasis in that it is concerned with extracting a maximum of information from an existing set of structure factors. The underlying unity of Wiener methods and information theory is readily developed along two distinct lines.

A. Filters. Shannon's Theorem 14 (Shannon & Weaver, 1949) requires that a signal passed through a linear filter with Fourier transform $X(f)$ undergo a gain in entropy proportional to

$$\int_W \ln |X(f)|^2 df, \quad (39)$$

where W indicates the frequency band to which components of $X(f)$ are limited. If an uncorrelated or white signal is passed through the filter then its entropy gain is given by (39), within a constant. A maximum of entropy gain may be sought for the filtering process, subject to suitable constraints, and the resulting $|X(f)|^2$ is the maximum entropy spectral estimate for the time series x (Ables, 1974; Robinson & Treitel, 1980). The constraints to be imposed are in the form of known ϕ , the autocorrelation of x ,

$$\int_W \Phi(f) \exp \{-2\pi i f k\} = \phi_k; -m \leq k \leq m, \quad (40)$$

and the expression to be maximized is

$$\int_W \left\langle \ln \Phi(f) - \sum_{k=-m}^m \lambda_k [\Phi(f) \exp \{-2\pi i f k\} - \phi_k] \right\rangle df, \quad (41)$$

in which Lagrange multipliers γ have been introduced.

Maximization of (41) results in

$$\Phi(f) = \frac{1}{\sum_{k=-m}^m \lambda_k \exp \{-2\pi i f k\}}. \quad (42)$$

It is clear that $\Phi(f)$ has been required to be positive definite and that this equation which gives a form for the maximum-entropy estimate of $\Phi(f)$ or $|X(f)|^2$ therefore may be written as

$$|X(f)|^2 = \frac{1}{|A(f)|^2} = \frac{\sigma_n^2/L}{\left| \sum_{t=0}^n \gamma_t \exp \{-2\pi itf\} \right|^2}. \quad (43)$$

Of course the foregoing right-hand expression has been arranged to correspond to equation (20) and to show that c_{ρ_x} in (25) is a maximum-entropy electron-density estimate. In regard to c_{ρ_x} , constraints given by (40) are the known structure factors.

B. *Density maps.* Gull and Daniell (1978) have proposed a maximum-entropy electron-density estimate based upon constrained maximization of

$$-\sum_x \rho_x \ln \rho_x, \quad (44)$$

the configuration entropy of the density function. As in the preceding case the constraints are the known structure factors. Gull and Daniell's formula for ρ is

$$\rho_x = \exp \left\langle -1 + \lambda \sum_{k \in K} \frac{1}{\sigma_k^2} \left[F_k^{(0)} - F_k^{(c)} \right] \right. \\ \left. \exp \{-2\pi i k x\} \right\rangle, \quad (45)$$

in which λ is a positive constant, K represents the set of observations, $F^{(0)}$ the observed structure-factor data with presumed error-free phases, σ_k the standard error for $F^{(0)}$, and $F^{(c)}$ the Fourier transform of the most recent iterate of ρ_x .

The formulations of density based upon maximizing of either $\Sigma \ln \rho$ or $-\Sigma \rho \ln \rho$ are obviously different. It should be observed, however, that in both cases the same structure-factor data would be used as constraints and in both cases ρ is required to be positive

definite. This latter requirement, in principle, demands a unique positive-definite estimate of ρ for fixed structure moduli and associated phases. In principle, use of either $\Sigma \ln \rho$ or $-\Sigma \rho \ln \rho$ ought to yield identical results except for the numerical effects of the manner in which structure factors enter the computations.

The differences between (25) and (45) are most substantial. Nevertheless, two particular similarities are important to crystallographic application. Both formulations for ρ provide a means of distributing error in a structure factor over the entire set, and both can yield true superresolution or resolution enhancement in a maximum-entropy electron-density estimate.

6. Conclusions

Wiener methods bring analytic-function theory to bear on computations involving electron density. Application of the methods turns upon the ideal positive-definite character of density functions and involves the determination of their analytic square roots. It is assumed that structure factors are available and free of such systematic errors as anomalous dispersion. Determination of an analytic root of electron density can also be considered deconvolution of structure factors and although Wiener methods require both, the operations need not be explicit.

Actual calculations emphasize the analytic-root aspect of the methods inasmuch as C_ρ is given as proportional to the squared modulus of an analytic function. Resolution enhancement, the initial goal of this work, is more readily related to the deconvolution operation in reciprocal space. Inasmuch as a root of C_ρ has a transform not limited in resolution by the experimental range of structure-factor data, the implied convolution whose transform is C_ρ is also potentially of unlimited extent in reciprocal space. Because structure-factor deconvolution is only implied,

resolution-enhancement factors have been estimated by comparison of density functions. It seems clear that the Wiener methods are most successful in one dimension for which resolution-enhancement factors of >2 may be expected for the qualitative resolution of individual atom profiles in an electron-density function.

In two dimensions, which illustrates n -dimensional generalization, resolution-enhancement factors up to 1.3 were found by empirical testing. The highest resolution factors were found, however, by forcing a matrix of structure factors toward singularity. Although the results clearly show resolution enhancement, the techniques used to maximize resolution are not suitable to general use. In n -dimensions it appears that inherently iterative schemes will be required for routine exploitation of resolution enhancement by Wiener methods.

The resolution enhancement of Wiener methods is the same phenomenon as the superresolution which accompanies formation of maximum-entropy images by information-theory methods. This is required by the form of C_ρ , which is exactly that for a maximum-entropy estimate of ρ , ρ treated as the squared modulus of the transform of a linear filter.

This work has been supported in part by the Robert A. Welch Foundation through grant A-742 and by the Research Corporation through a Cottrell Research Grant.

References

- ABLES, J. G. (1974). *Astron. Astrophys. Suppl.*, **15**, 383-393.
- BENJAMIN, R. (1980). *IEE Proc.* **F127**, 341-353.
- BROCKWAY, L. O. & ROBERTSON, J. M. (1939). *J. Chem. Soc.*, 1324-1332.
- CARATHÉODORY, C. (1964) *Theory of functions*, Vol. 1, New York: Chelsea.
- CLAERBOUT, J. F. (1976). *Fundamentals of Geophysical Data Processing*. New York: McGraw-Hill.
- COLLINS, D. M. (1978). *Acta Cryst.* **A34**, 533-541.

- Encyclopedic Dictionary of Mathematics* (1977). Cambridge, Massachusetts: MIT Press, 1049.
- GASSMANN, J. (1977). *Acta Cryst.* A33, 474-479.
- Gentleman, W. M. & SANDE, G. (1966). *AFIPS Proc.*, 29, 563-578.
- GOEDKOOP, J. A. (1950). *Acta Cryst.* 3, 374-378.
- GULL, S. F. & DANIELL, G. J. (1978). *Nature* (London), 272, 686-690.
- HELSON, H. & LOWDENSLAGER, D. (1958). *Acta Math.*, 99, 165-202.
- KARLE, J. & HAUPTMAN, H. (1950). *Acta Cryst.*, 3, 181-187.
- OPPENHEIM, A. V. & SCHAEFER, R. W. (1975). *Digital Signal Processing*. Englewood Cliffs, New Jersey: Prentice-Hall, Ch. 11.
- ROBINSON, E. A. (1967). *Statistical communication and Detection*. New York: Hafner.
- ROBINSON, E. A. & TREITEL, S. (1980). *Geophysical Signal Analysis*. Englewood Cliffs, New Jersey: Prentice-Hall.
- SHANNON, C. E. & WEAVER, W. (1949). *The Mathematical Theory of Communication*. Urbana, Illinois: University of Illinois Press.
- WIENER, N. (1949). *Extrapolation, Interpolation, and Smoothing of Stationary Time Series*. New York: Wiley.

SUBJECT INDEX

Note. Certain frequently occurring expressions, mainly statistical, with or without qualifying adjectives, are not exhaustively indexed. These include Atomic scattering factor, Correlation, Covariance, Deviation, Distribution, Mean Parameter, Space group, Statistics, Structure factor, Variance, and some others.

- Absolute configuration, 133, 134
- Absolute intensities, from relative, 1–2
- Absorption, as source of bias, 234
 - edge, 134
- Acentric distribution, 5–8, 25, 73–74, 102, 123–125, 142
 - variance of, 8
 - see also* Space group *P1*
- N-Acetyl-allo-hydroxy-L-proline lactone, 215
- Alternatives to least squares, 3, 4, 225–226, 229–262, 273–298 269–272
 - see also* Least squares
- Alternatives to *R* tests, 19, 46, 48, 187–193
- Ammonium hydrogen malate, 203, 220–223
- Anomalous dispersion, Anomalous scattering, 133–172
 - see also* Dispersion
- Argand diagram, 137
- Arvesen's test, 187, 189–193
- Atomic heterogeneity, *see* Heavy atom
- Atomic scattering factor, 301 (note)
 - affected by dispersion, 135–136
- Autocorrelation, 276 *et seq.*

- Background, important in determining accuracy of intensity estimation, 33–37, 181–182
- Bayes' theorem, 10, 22–23, 213
- Bayesian statistics, 3, 10, 19–51
- Bessel function, modified, 12, 103, 123, 201, 213
- Bias, in density estimation, 182, 226
 - in parameter estimation, 4, 225, 230, 254, 256
- Bicentric distribution, 127–128
- Bijvoet differences, Bijvoet pairs, 29, 133–171
- Bijvoet ratio, defined, 138
- Biweight function, 237, 240, 242, 245

- Cauchy distribution, 109–110, 237–238
 - contamination of normal, 190, 193

- Central-limit theorem
 - independent variables, 2, 57–58, 60, 84, 102–103
 - non-independent variables, 176
 - Centric distribution, 8–9, 25, 73–74, 102, 119–123
 - variance of, 9
 - see also* Space group $P\bar{1}$
 - Centroid method of locating reflexion, 267
 - contrasted with profile fitting, 40–44
 - Chi-square agreement statistic, 256
 - Communication engineering, 274
 - Computer programs, INSTAT, 95
 - mean values of powers of structure factors, 71–73
 - MULTAN, 95
 - NORMAL, 95
 - RFINE 4, 245, 247, 248
 - Computer simulation, of error distributions, 190–193
 - of intensity distributions, 53, 55–56
 - of non-ideal distributions, 58–66
 - of R_s distributions, 207–208
 - Constraints, refinement under, 19, 44–49
 - Convergence, improvement of, 248, 261
 - Correlation coefficient, 13–14, 301 (note)
 - autocorrelation, 276 *et seq.*
 - Correlation of atomic positions, 4, 175–177
 - of parameters, 230
 - Counting statistics, 9–10, 24, 39, 301 (note)
 - as recorded, 179–185
 - assumed to have a normal distribution, 26–27
 - Covariance, 13, 188, 268, 301 (note)
 - Crystallographic statistics, 301 (note)
 - Bayesian, 10–11, 19–51
 - general review, 5–17, 53–55, 83–92
 - heavy-atom dependence, 53–81, 105–110, 117–131
 - non-ideal, 53–81, 83–97, 175–177
 - origin of, 1–2
 - space-group dependence, 53–96, 155–157
 - Cumulants, 83, 86–92, 93–94
- .
- Data reduction, 267–268
 - Deformation density, Deformation potential, 179
 - Dimers, test of symmetry of, 46
 - cis*, *cis*-4, 6-Dimethyl trimethylenesulphite, 205
 - Direct methods, *see* Structure determination
 - Discrepancy index (R factor), 14, 19, 46, 48, 59, 60, 63, 195–223, 232
 - moments of, 209–219
 - Discriminator criterion, 195–223

- Disordered structures, 2
- Dispersion (including anomalous dispersion, scattering), 3, 16, 29, 56, 69, 85, 133–172
- Edgeworth expansion, 53, 56, 78–81, 93, 119–125
 - and non-independence, 175–177
 - convergence of, 75–80
 - in phase determination, 99, 101, 104, 107, 110, 112–113
- Electron density
 - by dispersion, 15
 - by Fourier synthesis, 15, 179, 225–226, 265–266
 - by heavy-atom method, 16
 - maximum-entropy estimate, 296–297
 - reduction of error in, 267–268
 - Wiener methods for, 3, 273–298
- L-Ephedrine hydrochloride, 170
- Equivalent reflexions, 233 *et seq.*
- Errors, analysis of, 267–268
 - distribution function for, 190, 193, 226, 230
 - of model, 30–32, 259
 - of observation, 32, 39, 270
 - random (instrumental instability), 40, 184–185
 - statistical fluctuation, 9–10, 15, 179–185, 265–266
 - systematic, 15, 180, 244, 255–262
- Excess, Kurtosis, 37, 215
- Expected value, 27, 34, 38, 301 (note)
 - see also* Intensity of reflexion, mean value of
- Extinction, 245–261 *passim*
- Extinction parameter, defined, 245
- Extrapolation, 295, 296
- Filters, in extrapolation of series, 279–287
 - in information theory, 295–296
- Fixed-count timing, 9
- Fixed-time counting, 9
- Forsterite, 182–184
- Fourier transforms, as basis of Wiener methods, 274 *et seq.*
- Frequentist statistics, contrasted with Bayesian, 19–20
- Friedel's law, 134
 - see also* Dispersion
- Gamma distribution, 190
 - function, incomplete, 107
- Gaussian distribution, *see* Normal distribution
- Generalized distributions, 53–132
- 'Globs' (Harker), 176

304 *Subject Index*

- Gram-Charlier expansions, 53–54, 56, 78–81, 93, 119–125
 and non-independence, 175–177
 convergence of, 75–80
 in phase determination, 99–101, 103–104, 110
- Hamilton's R test, *see* R test
- Heavy atom, and dispersion, 133
 effect on intensity distributions, 55–66, 92, 105–110, 117–131
- Heavy-atom analysis, 195
- Heavy-atom method, *see* Structure determination
- Hermite polynomials, 68, 73–74, 75, 89, 120
- Hexamethylbenzene, 292–294
- Histograms, of R_2 distributions, 220
 of structure-factor distributions, 60–62, 76
 stem-and-leaf display, 257–259
- Hydrogen parameters, 245–262 *passim*, 301 (note)
- Hypercentric distribution, 126–128, 131
- Hypergeometric function, 214
- Hypersymmetry, 92, 131
- Hypothesis testing, 44–49, 187–193
- Information theory, 294–297
- INSTAT, 95
- Instrument instability, 39, 40, 41, 185
- Instrumental variance, 268
- Insulin, 40–42
- Intensity of reflexion, 301 (note)
 effect of correlation of atomic positions on, 4, 175–177
 mean value of, 2, 8, 24, 27, 176–177
 measurement of, 33–44, 179–185
 model for, 33–36
 probability distribution of large intensities, 4
 of small and moderate intensities, 2–3, 5–131; *see* more
 specific entries (acentric, bicentric, centric, *etc.*)
 symmetry dependence of, 54, 60–66, 70–77, 84–96, 117–131
 variance of, 9, 27, 180–182
- Intensity statistics, 301 (note)
 and non-independence, 175–176
 of recorded counts, 179–185
 reviews, 5–17, 53–97
- Interpolation, 273, 274
- Isomorphism, 11–14, 15–16
- Jackknife test, 187–193
- Jacobian, 28
- Karle-Hauptman determinants, 273, 283
- Kurtosis, Excess, 37, 215

- Laguerre polynomials, 68, 73–74, 75
 Least-squares refinement, 15, 27, 225–226, 269–272
 see also Alternatives to least squares
 Likelihood, 23, 31, 225–226
 Lindeberg-Lévy central-limit theorem, 57, 60
 L-Lysine hydrochloride dihydrate, 170
- Macromolecular crystals, 150–151, 273
 see also Protein crystallography
 MAD, defined, 240
 Markov chain, 114
 Matrix equations, solution of, 281–282
 Maximum-entropy estimates, 295–297
 Maximum likelihood, 225–226
 Mean intensity, *see* Intensity of reflexion
 Measurability of Bijvoet differences, 133–172
 Measurement of intensity, 9–11, 33–44, 179–185
 Median absolute deviation, 240–241
 Minimum-delay operator, 282–283
 Model, 20, 301 (note)
 Bayesian three-stage, 19, 29–33
 structural, 14, 46, 196
 Modified Bessel function, 12, 103, 123, 201, 213
 Monochromatization, 182
 MULTAN, 95
 Multiple-counter techniques, 42–43
- ‘Negative’ intensities, 10, 24–25, 182
 Non-crystallographic symmetry, 46, 126–128, 131
 neglect of, 56
 partial centrosymmetry, 133, 158–164
 Non-ideal distributions, 53–132
 NORMAL (program), 95
 Normal distribution, Gaussian distribution, 99, 101–112, 265
 301 (note)
 approximation to distribution of R_2 , 211, 215
 in significance testing, 190–193
 inadequate for large deviations, 226, 230, 254
 of measurements of an intensity, 26–27
 of structure factors of a centrosymmetric crystal, 8–9
 see also Centric distribution
 Normalized Bijvoet difference, defined, 137
 Normalized structure factor, 89, 100–101, 113, 138, 301 (note)
- Odd moments, vanishing of, 88
 Ordinate analysis of intensity measurement,
 contrasted with profile fitting, 40–43
 C. S.—20

- Orthogonal polynomials, 67–68, 175–177
see also Hermite polynomials, Laguerre polynomials
- Parameter estimation, 15, 20, 21, 25, 31, 36, 301 (note)
 bias in, 4, 225, 230, 254, 256
- Partial symmetry, 96, 133, 139, 158–164
- Patterson coefficient, 275
- Patterson function, 15, 114, 164, 225–226
- Peak position, 36, 267
- Peak shape, model for, 34–36
- Peak-to-background ratio, 180–182, 295
- Peak width, 36, 267
- Phase determination, 131
 errors in, 16
 from Bijvoet differences, 133–172
 Probability of validity, 99–114
- Phase shift on scattering, 134
- Pitman's test, 187, 190–193
- Point groups, 72, 124–125
- Poisson distribution, approximated by normal, 39
 variance of, 268
- Posterior distribution, 10, 23, 27, 28–29 (figure),
- Prealbumin, 43–45
- Prediction of stationary series, 276 *et seq.*
- Prior distribution, 10, 22, 26, 28–29 (figure)
- Profile fitting, 19, 33–44
- Protein crystallography, 14, 29, 33, 40–46 *passim*, 58, 150–151, 176, 273
- Pseudo-observations, defined, 44
- Pseudosymmetry, *see* non-crystallographic symmetry, Partial symmetry
- Pyrene, 127
-
- R* factor, *see* Discrepancy index
- R* test, 19, 46, 48, 187–193
- R_2 , moments of, 209–219
- Rayleigh distribution, 101, 110, 113–114
- Recorded counts, statistics of, 10, 33–45, 179–195
- Residual, *see* Discrepancy index
- Resistant, *see* Robust, Robustness
- Resolution enhancement, 273–298 *passim*, specifically 286–287, 294, 297–298
- Restrained refinement, 44–46
- Robust/Resistant techniques, 229–262
- Robustness, 193, 226, 229–262
- Rotation search, 195, 219–223
- Rubidium di-*o*-nitrobenzoate, 125

- Scale factor, 225, 245, 246, 268
- Scientific method, 24, 47–48, 269–272
- Secondary extinction, 245
- Secondary minima, 269–272
 - see also* Series termination
- Semi-invariants (Cumulants), 83, 86–92, 93–94
- Seminvariants (Structure seminvariants), 117, 118, 128–131
- Series termination, 265–266, 273–298
- Single Crystal Intensity Project, 229–231, 243–262
- Skewness, 37, 211–212, 215, 216
- Smoothing, 265, 273
- Soft constraints, 19, 44–49
- Space groups, 301 (note)
 - Fdd2*, *Fddd*, 72, 94
 - of higher symmetry, 26, 69–75, 88, 94, 155–157,
 - P1*, 2, 5–8, 25, 80, 117, 139, 142–157, 197–207
 - P1̄*, 8–9, 26, 57, 59–61, 63–66, 76–77, 117, 197–207
 - P2₁*, 245
 - Pmmm*, 57, 60, 62, 77
- Spectral estimation, 278
- Standard deviation, *see* Variance
- Standardized moments, 93–94
- Stationary series, 273, 274
- Stem-and-leaf display, 257–259
- Stereochemistry, effect on intensity statistics, 175–177
- Structural isomorphism, 11–14, 15–16
- Structure determination, 301 (note)
 - bias in, 3, 182
 - direct methods, 3, 16, 113, 131, 195
 - dispersion, 16, 133–172
 - heavy-atom method, 195
 - isomorphous replacement, 11–17
 - least squares, 15, 269–272
 - robust/resistant technique, 229–262
 - tangent formula, 16–17
 - use of residual R_s , 195–223
 - see also* Electron density, Parameter estimation
- Structure, factor, trigonometric, 57–58, 68–73, 83, 86–88, 155
- Structure factors, probability distribution of, *see* Crystallographic statistics, Intensity of reflexion
- Structure seminvariants, 117, 118, 128–131
- Subjectivity, recognition as an advantage, 21–22, 48–49
- Systematic errors, 15, 180, 244, 255–262

- Tangent formula, 16–17
- D-Tartaric acid, 229, 231, 243–262
- Temperature parameters, 225, 245–262

308 *Subject Index*

7, 7, 8, 8-Tetracyanoquino-dimethane-phenazine complex, 127-128

Thermal parameters, 225, 245-262

Time series, 274

Translation search, 195

Truncation of weak intensities, 140, 145-148, 171, 196-223

L-Tyrosine hydrochloride, 170

Unobserved reflexions, 7, 8, 9-10, 140, 171, 182

Variance (including some references to standard deviation),
8, 9, 14, 27, 36, 37, 38, 39, 232-262 *passim*, 282, 268

Weighting, bias from 4, 182, 225-226

in alternatives to *R* tests, 188

in electron-density synthesis, 15-16

in least squares, 15, 225-226, 232-252

Width of peak, 36, 267

Wiener methods, 273-298

Wilson distributions, Wilson statistics, 5-9, 25, 48, 57, 59, 84,
100, 117, 142, 198

Window, determination of setting of, 40-44

AUTHOR INDEX

Note: The page numbers in this index are normally those on which the article containing the authors cited begins, and not the page on which the name actually occurs.

Ables, J. G., 273
Abrahams, S. C., 133, 229
Abramowitz, M., 53, 83, 195
Andrews, D. F., 229
Andrews, L. C., 19, 267
Arsenin, V. Ja, 265
Arvesen, J. N., 187

Banner, D. W., 19
Bard, Y., 225
Barnett, V., 19
Beaton, A. E., 229
Beevers, C. A., 229
Belgaumkar, J., 99
Bell, W. D., 19, 187
Benjamin, R., 273
Bernstein, S., 175
Bertaut, E. F., 83, 99
Beu, K. E., 225
Bijvoet, J. M., 133
Blake, C. C. F., 19
Box, G. E. P., 19
Brockway, L. O., 273
Broyden, C. G., 229
Burridge, J. M., 19
Busing, W. R., 229

Caratheodory, C., 273
Clarebourt, J. F., 273
Cochran, W., 1, 99
Collins, D. M., 273
Collin, R. L., 99, 117
Cook, R. D., 229
Cox, G. S., 229
Cox, J. M., 19
Cramér, H., 53, 83, 117
Critchley, S. R., 19
Cromer, D. T., 229

Dale, D., 133
Daniel, G. J., 273

310 *Author Index*

- Davis, C. L., 225
Declercq, J. P., 83
DeTinetti, B., 19
DeGroot, M. H., 19
De Polignac, C., 19
Dodson, E. J., 19
Dodson, G. G., 19
Donohue, J., 99
- Edwards, A. W. F., 225
Evans, P. R., 19
- Faggiani, R., 53
Feller, W., 7, 99
Finger, L. W., 229
Foster, F., 83, 117, 133
French, S., 7, 19, 99, 175
- Gabe, E. J., 267
Garfield, E., 1
Gassmann, J., 273
Gauss, C. F., 269
Geise, H. J., 195
Geisow, M. J., 19
Gentleman, W. M., 273
Germin, G., 83
Giacovazzo, C., 1, 99, 117
Goedkoop, J. A., 273
Goldberg, I., 83, 117
Grant, D. F., 267
Gull, S. F., 273
- Hall, G., 117, 133
Hamilton, W. C., 19, 187, 225, 229
Harker, D., 1, 175
Hargreaves, A., 83, 99, 117, 133
Hauptman, H., 1, 7, 19, 53, 83, 99, 117, 133, 273
Helmoldt, R. B., 179
Helson, H., 273
Hodgkin, D. C., 19, 133
Howells, E. R., 99, 117
Huber, P. J., 229
Hughes, E. W., 1
Hull, S. C., 83
- James, R. W., 133
Jeffreys, H., 19
Jia-Xing, Yao, 7

Johnson, N. J., 187

Johnson, N. L., 19

Kaldor, U., 53, 83, 133

Karle, J., 1, 19, 53, 83, 99, 117, 133, 273

Kay, M. I., 229

Kendall, M. G., 83

Klug, A., 83, 99, 117

Kuh, E., 229

Lantsosh, K., 265

Larsen, F. K., 267

Lattman, E. E., 195

Lehman, M. S., 267

Lempers, F. B., 19

Lenstra, A. T. H., 195

Lessinger, L., 83

Levy, H. A., 229

Lewis, M. L., 19

Lewitova, A., 19

Li, W. K., 187

Lindley, D. V., 19

Lippert, B., 53

Lipson, H., 1, 117

Lock, C. J. L., 53

Lowdenslager, D., 273

Mackenzie, J. K., 229

Main, P., 83

Mallows, C. L., 229

Mandel, J., 225

Mann, J. B., 229

Marsh, D. J., 19

Martin, K. O., 229

Mathieson, A. McL., 229

Maslen, E. N., 133, 225

Maslen, E. R., 1, 117

McCandlish, L. E., 19, 267

Mendes, M., 19

Mitra, G. B., 99

Muirhead, H., 19

Musil, F. J., 225

Nelmes, R. J., 267

Nicholson, W. L., 229

Nielson, K., 225

Nigam, G. D., 117

312 *Author Index*

- Oatley, S. J., 19
Okay, Y., 229
Oppenheim, A. V., 273
- Parthasarathy, S., 1, 7, 19, 53, 83, 117, 133, 195
Parthasarathi, V., 133, 195
Peerdeman, A. F., 133
Petit, G. H., 195
Phillips, D. C., 19, 99, 117
Pitman, E. J. G., 187
Ponnuswamy, M. N., 133
Price, P. F., 225
Prince, E., 229
- Ramachandran, G. N., 133
Raman, S., 133
Ramaseshan, S., 1, 133
Rees, B., 179, 225
Rérat, B., 19
Rérat, C., 19
Reynolds, C. D., 19
Richardson, M. F., 1, 19, 187
Rigoult, J., 267
Robertson, J. M., 273
Robinson, E. A., 273
Rogers, D., 19, 99, 117
Rossman, M. G., 195
Rothstein, S. M., 19, 187
- Sabesan, M., 19
Sabine, T. M., 229
Sande, G., 273
Sass, R. L., 99
Sayre, D., 1, 99
Schafer, R. W., 273
Shannon, C. E., 273
Shevyrev, A. A., 265
Shmueli, U., 53, 83, 99, 117, 133, 175
Shotton, D. M., 19
Sim G. A., 99, 117, 133
Simonov, V. I., 265
Smith, A. F. M., 19
Spiegelhalter, D. J., 19
Srinivasan, R., 1, 7, 19, 53, 83, 117, 133, 195
Stegun, I. A., 53, 83, 195
Stemple, N. R., 229
Stern, F., 229

Stout, G. H., 19, 267
Stuart, A., 83
Swaminathan, P., 133

Taio, G. C., 19
Taylor, G. H., 229
Tickle, I., 19
Tikhonov, A. N., 265
Tollin, P., 195
Treitel, S., 273
Tukey, J. W., 229

van Bommel A. J., 133
van De Mieroop, W., 195
van Havere, W., 195
van Loock, J. F., 195
Varghese, J. N., 225
Velmurugan, D., 133
Venkatesan, K., 133
Versichel, W., 195
Vijayalakshmi, B. K., 133
Vos, A., 179

Watson, G. N., 7
Watson, H. C., 19
Weaver, W., 273
Welsch, R. E., 229
Whitney, D. R., 225
Wiener, N., 273
Wilson, A. J. C., 1, 7, 53, 83, 99, 117, 133, 175, 179, 195, 225
Wilson, K., 7, 99, 175
Wilson, K. S., 19
Woolfson, M. M., 83, 99, 117

Yow-Lam Oh, 133
Yü, S. H., 1

Zachariasen W. H., 133, 229

